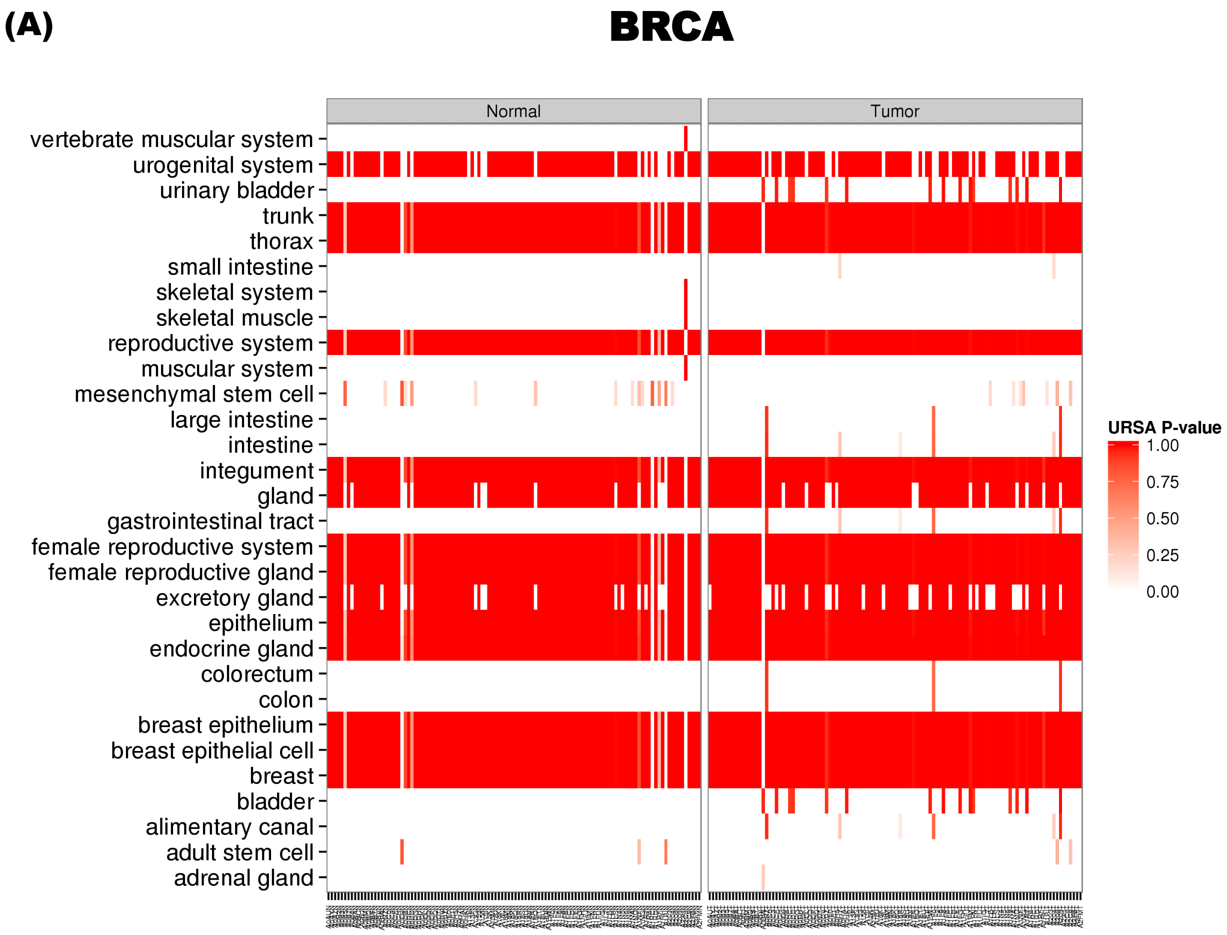
Supplementary Material

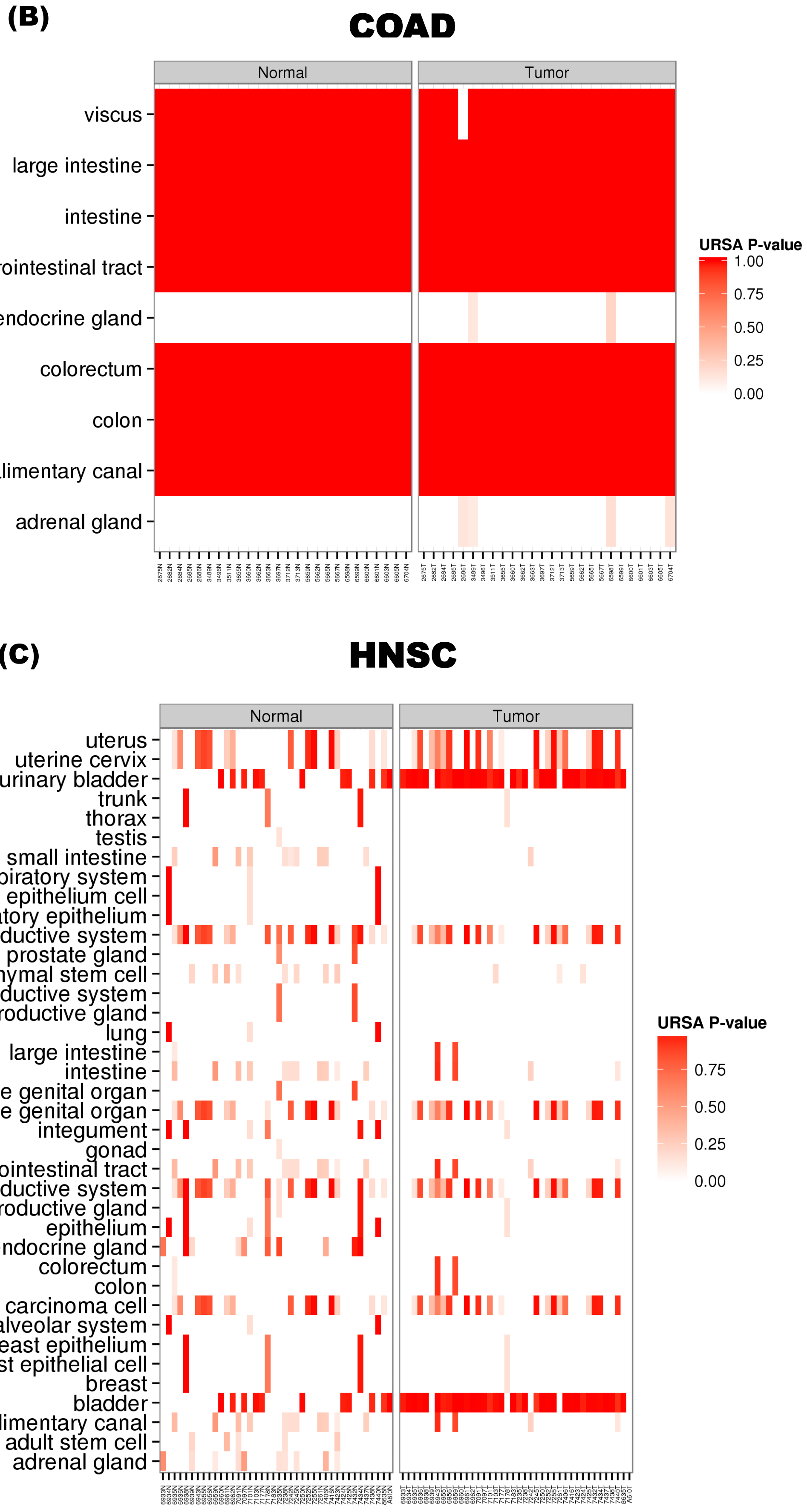
Recurrent alternative splicing isoform switches in tumor samples provide novel signatures of cancer

Endre Sebestyén[1], Michał Zawisza[2], Eduardo Eyras[1,3,4]

[1]Computational Genomics, Universitat Pompeu Fabra, Dr. Aiguader 88, E08003 Barcelona, Spain
[2]Universitat Politècnica de Catalunya, Jordi Girona 1-3, Barcelona E08034, Spain
[3]Catalan Institution for Research and Advanced Studies, Passeig Lluís Companys 23, E08010 Barcelona, Spain
[4]Correspondence to: eduardo.eyras@upf.edu

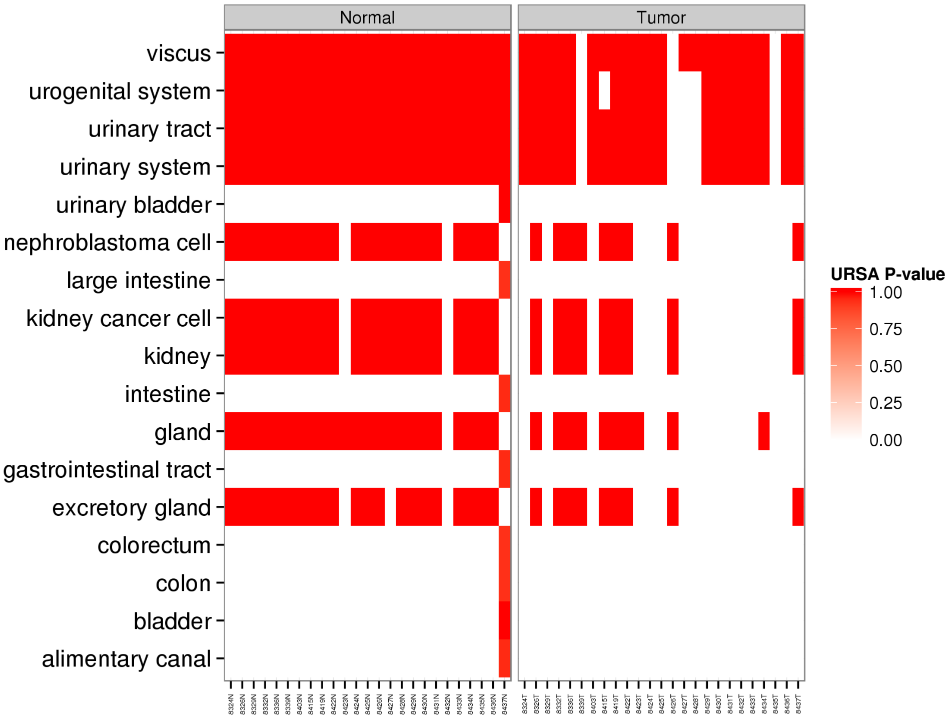


Supplementary Figure 1. For each cancer type, we run URSA (Lee et al. 2013) on the estimated read-counts per gene for all the paired-samples and kept only the tissue predictions with posterior probability P>0.1 (y axis). Tumor and normal samples (x axis) were then clustered separately. Both paired-samples from the same patient were removed if either of the two, tumor or normal, did not cluster with the rest of the same type. The heatmaps show the selected samples for each cancer type: BRCA **(A)**, COAD **(B)**, HNSC **(C)**, KICH **(D)**, KIRC **(E)**, LUAD **(F)**, LUSC **(G)**, PRAD **(H)** and THCA **(I)**.


(B)
COAD
Normal
Tumor
viscus
large intestine
intestine
gastrointestinal tract
endocrine gland
colorectum
colon
alimentary canal
adrenal gland
URSA P-value
1.00
0.75
0.50
0.25
0.00
2675N 2682N 2684N 2685N 2686N 3489N 3496N 3511N 3655N 3660N 3662N 3663N 3697N 3712N 3713N 5659N 5662N 5665N 5667N 6598N 6599N 6600N 6601N 6603N 6605N 6704N
2675T 2682T 2684T 2685T 2686T 3489T 3496T 3511T 3655T 3660T 3662T 3663T 3697T 3712T 3713T 5659T 5662T 5665T 5667T 6598T 6599T 6600T 6601T 6603T 6605T 6704T
(C)
HNSC
Normal
Tumor
uterus
uterine cervix
urinary bladder
trunk
thorax
testis
small intestine
respiratory system
respiratory epithelium cell
respiratory epithelium
reproductive system
prostate gland
mesenchymal stem cell
male reproductive system
male reproductive gland
lung
large intestine
intestine
internal male genital organ
internal female genital organ
integument
gonad
gastrointestinal tract
female reproductive system
female reproductive gland
epithelium
endocrine gland
colorectum
colon
cervical carcinoma cell
bronchoalveolar system
breast epithelium
breast epithelial cell
breast
bladder
alimentary canal
adult stem cell
adrenal gland
URSA P-value
0.75
0.50
0.25
0.00
6933N 6934N 6935N 6936N 6938N 6939N 6943N 6955N 6956N 6959N 6960N 6961N 6962N 7091N 7097N 7101N 7103N 7177N 7178N 7183N 7235N 7238N 7242N 7245N 7250N 7252N 7255N 7261N 7406N 7416N 7423N 7424N 7425N 7432N 7434N 7437N 7438N 7440N 8635N A6I0N
6933T 6934T 6935T 6936T 6938T 6939T 6943T 6955T 6956T 6959T 6960T 6961T 6962T 7091T 7097T 7101T 7103T 7177T 7178T 7183T 7235T 7238T 7242T 7245T 7250T 7252T 7255T 7261T 7406T 7416T 7423T 7424T 7425T 7432T 7434T 7437T 7438T 7440T 8635T A6I0T

KICH

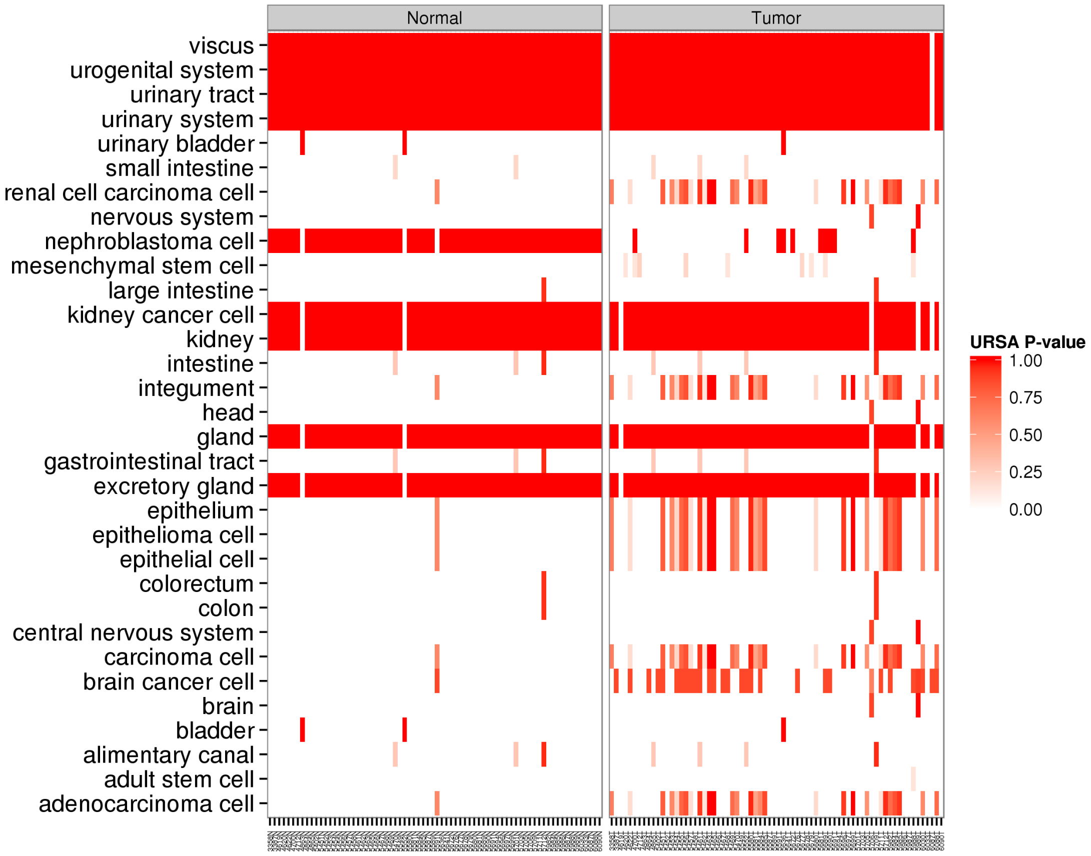
(D)



Normal
Tumor
viscus
urogenital system
urinary tract
urinary system
urinary bladder
nephroblastoma cell
large intestine
kidney cancer cell
kidney
intestine
gland
gastrointestinal tract
excretory gland
colorectum
colon
bladder
alimentary canal
URSA P-value
1.00
0.75
0.50
0.25
0.00
8324N
8326N
8329N
8332N
8336N
8339N
8403N
8415N
8419N
8422N
8423N
8424N
8425N
8426N
8427N
8428N
8429N
8430N
8431N
8432N
8433N
8434N
8435N
8436N
8437N
8324T
8326T
8329T
8332T
8336T
8339T
8403T
8415T
8419T
8422T
8423T
8424T
8425T
8426T
8427T
8428T
8429T
8430T
8431T
8432T
8433T
8434T
8435T
8436T
8437T

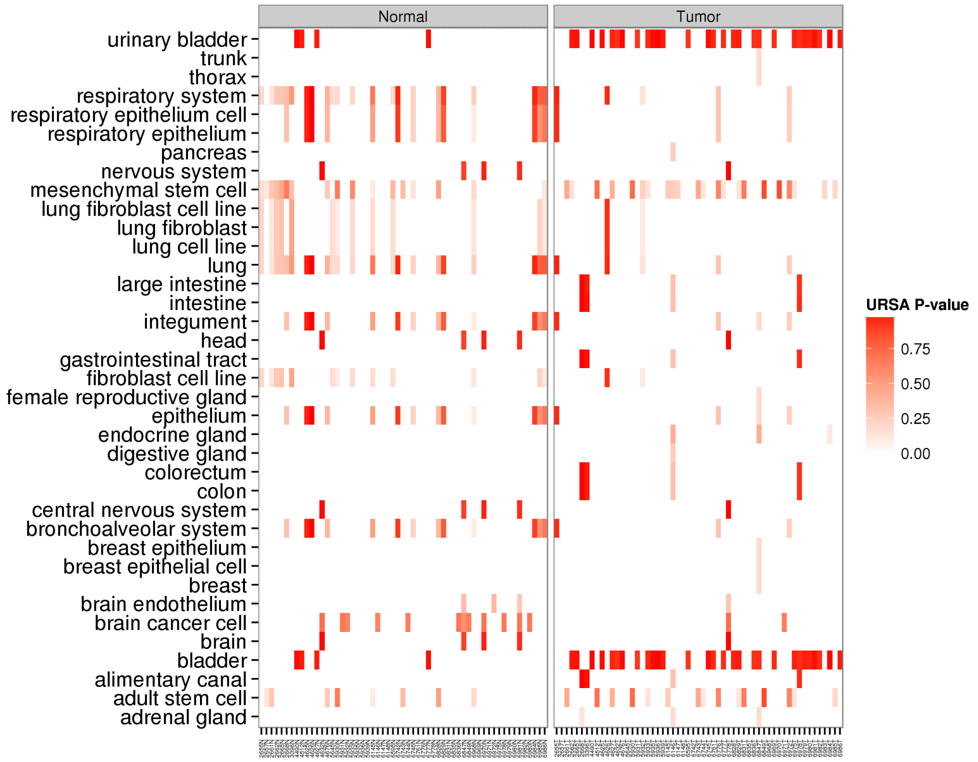
(E)

KIRC

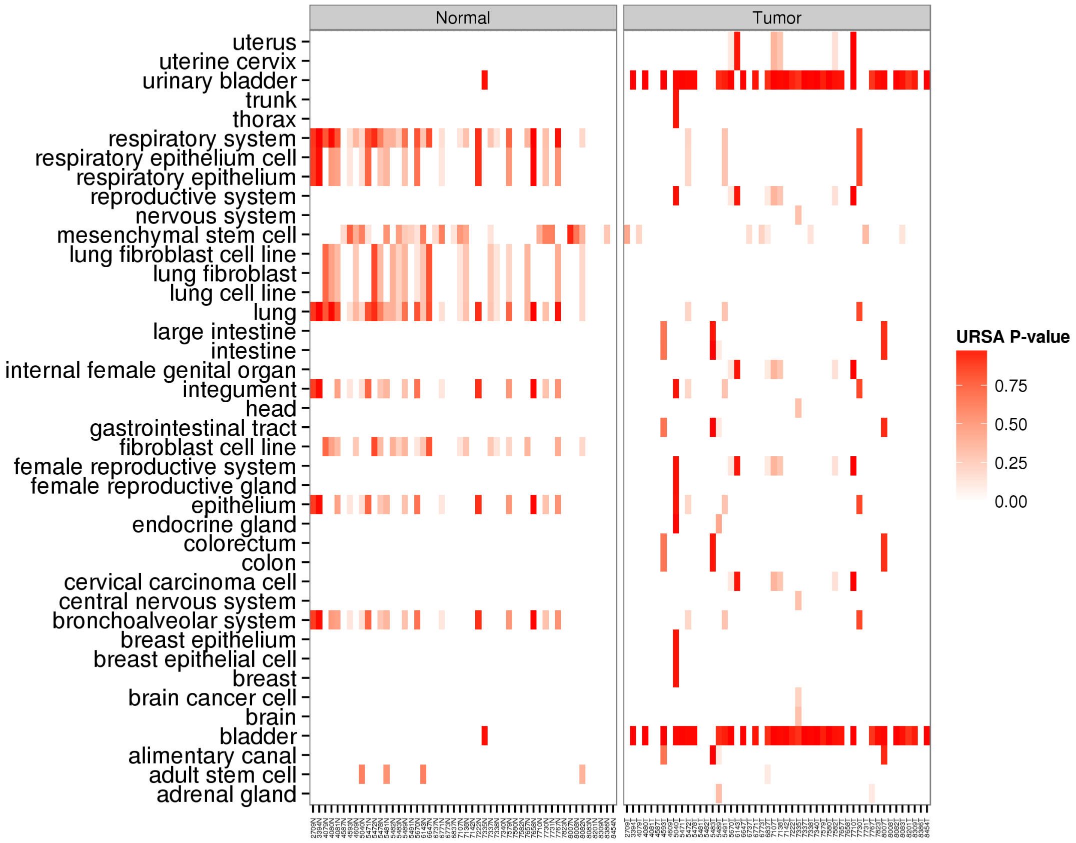


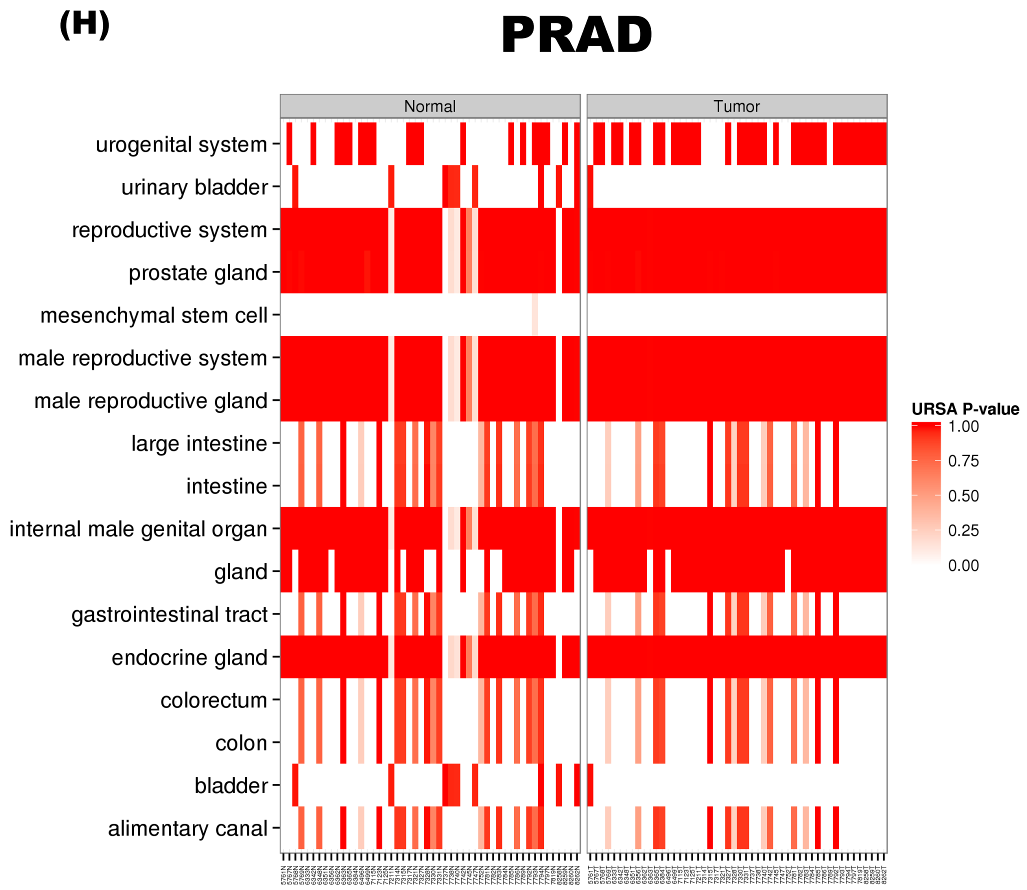
Normal
Tumor
viscus
urogenital system
urinary tract
urinary system
urinary bladder
small intestine
renal cell carcinoma cell
nervous system
nephroblastoma cell
mesenchymal stem cell
large intestine
kidney cancer cell
kidney
intestine
integument
head
gland
gastrointestinal tract
excretory gland
epithelium
epithelioma cell
epithelial cell
colorectum
colon
central nervous system
carcinoma cell
brain cancer cell
brain
bladder
alimentary canal
adult stem cell
adenocarcinoma cell
URSA P-value
1.00
0.75
0.50
0.25
0.00

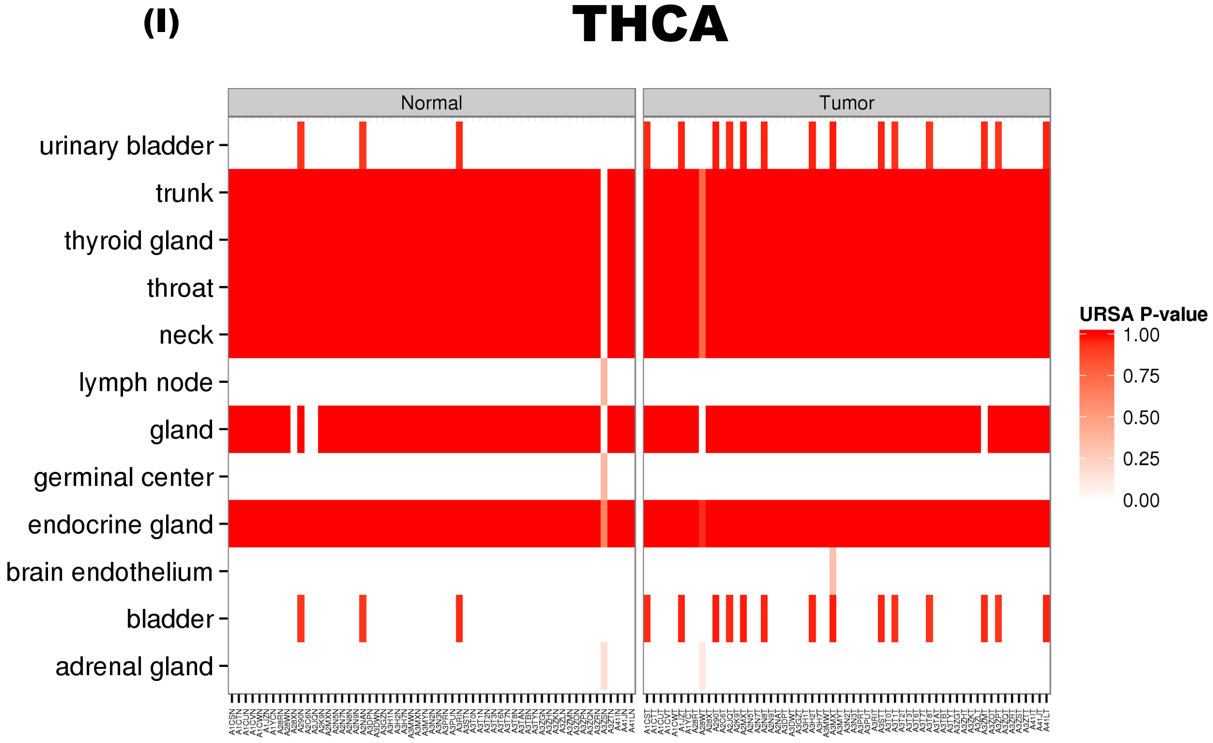
(F) LUAD

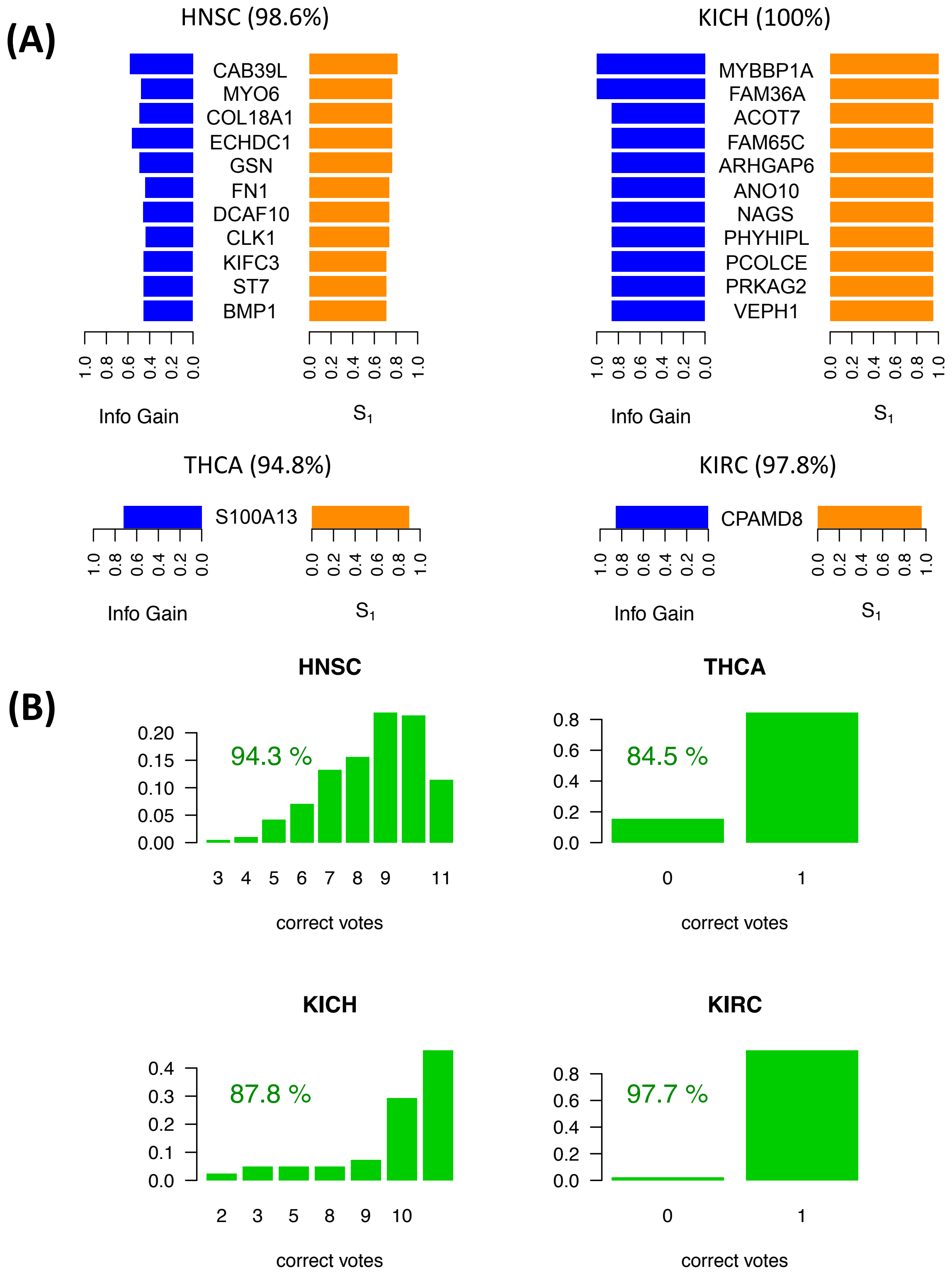

Normal
Tumor
urinary bladder
trunk
thorax
respiratory system
respiratory epithelium cell
respiratory epithelium
pancreas
nervous system
mesenchymal stem cell
lung fibroblast cell line
lung fibroblast
lung cell line
lung
large intestine
intestine
integument
head
gastrointestinal tract
fibroblast cell line
female reproductive gland
epithelium
endocrine gland
digestive gland
colorectum
colon
central nervous system
bronchoalveolar system
breast epithelium
breast epithelial cell
breast
brain endothelium
brain cancer cell
brain
bladder
alimentary canal
adult stem cell
adrenal gland
URSA P-value
0.75
0.50
0.25
0.00

(G) LUSC


Normal
Tumor
uterus
uterine cervix
urinary bladder
trunk
thorax
respiratory system
respiratory epithelium cell
respiratory epithelium
reproductive system
nervous system
mesenchymal stem cell
lung fibroblast cell line
lung fibroblast
lung cell line
lung
large intestine
intestine
internal female genital organ
integument
head
gastrointestinal tract
fibroblast cell line
female reproductive system
female reproductive gland
epithelium
endocrine gland
colorectum
colon
cervical carcinoma cell
central nervous system
bronchoalveolar system
breast epithelium
breast epithelial cell
breast
brain cancer cell
brain
bladder
alimentary canal
adult stem cell
adrenal gland
URSA P-value
0.75
0.50
0.25
0.00

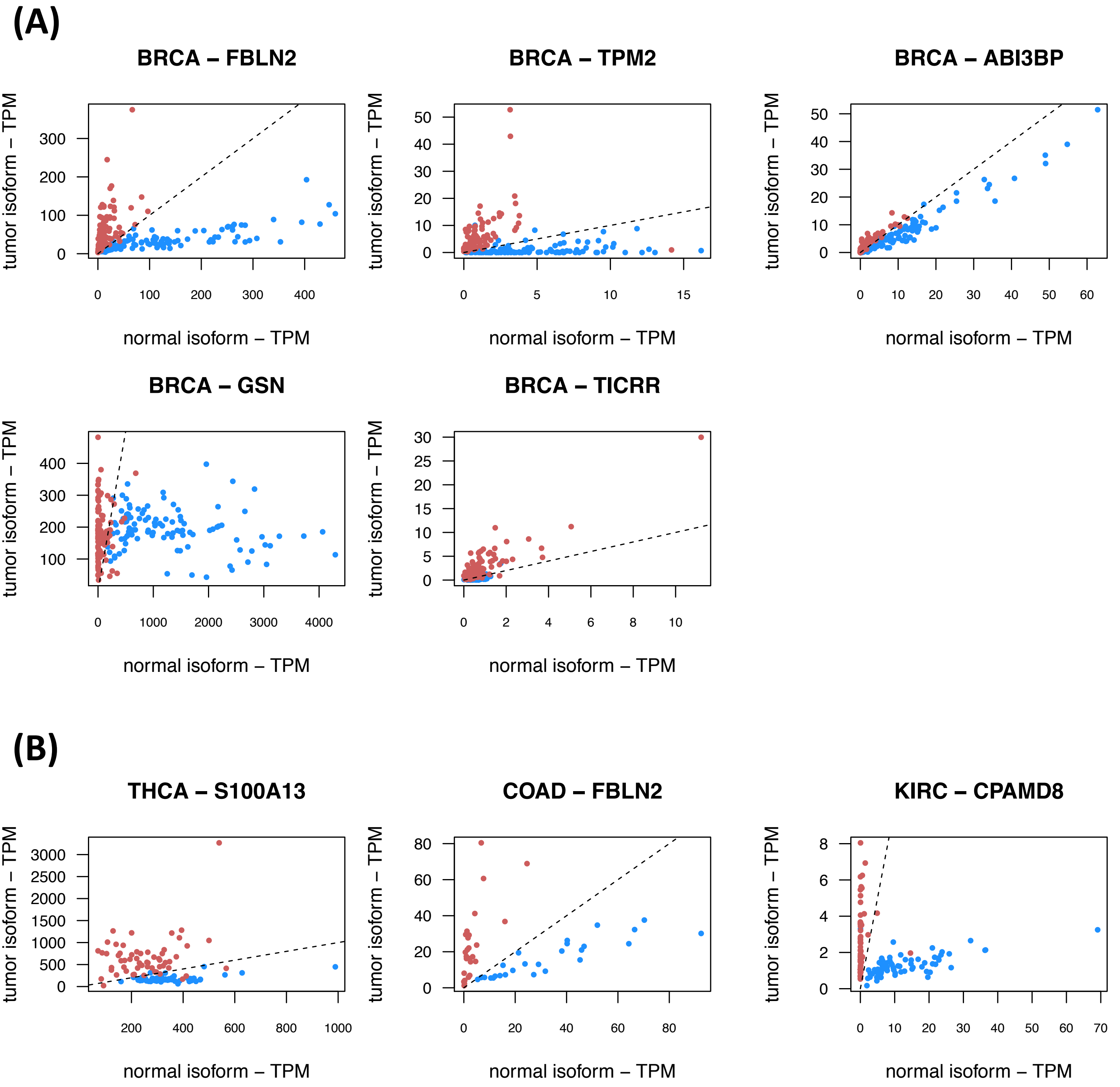

(H)
PRAD
Normal
Tumor
urogenital system
urinary bladder
reproductive system
prostate gland
mesenchymal stem cell
male reproductive system
male reproductive gland
large intestine
intestine
internal male genital organ
gland
gastrointestinal tract
endocrine gland
colorectum
colon
bladder
alimentary canal
URSA P-value
1.00
0.75
0.50
0.25
0.00


(I)
THCA
Normal
Tumor
urinary bladder
trunk
thyroid gland
throat
neck
lymph node
gland
germinal center
endocrine gland
brain endothelium
bladder
adrenal gland
URSA P-value
1.00
0.75
0.50
0.25
0.00

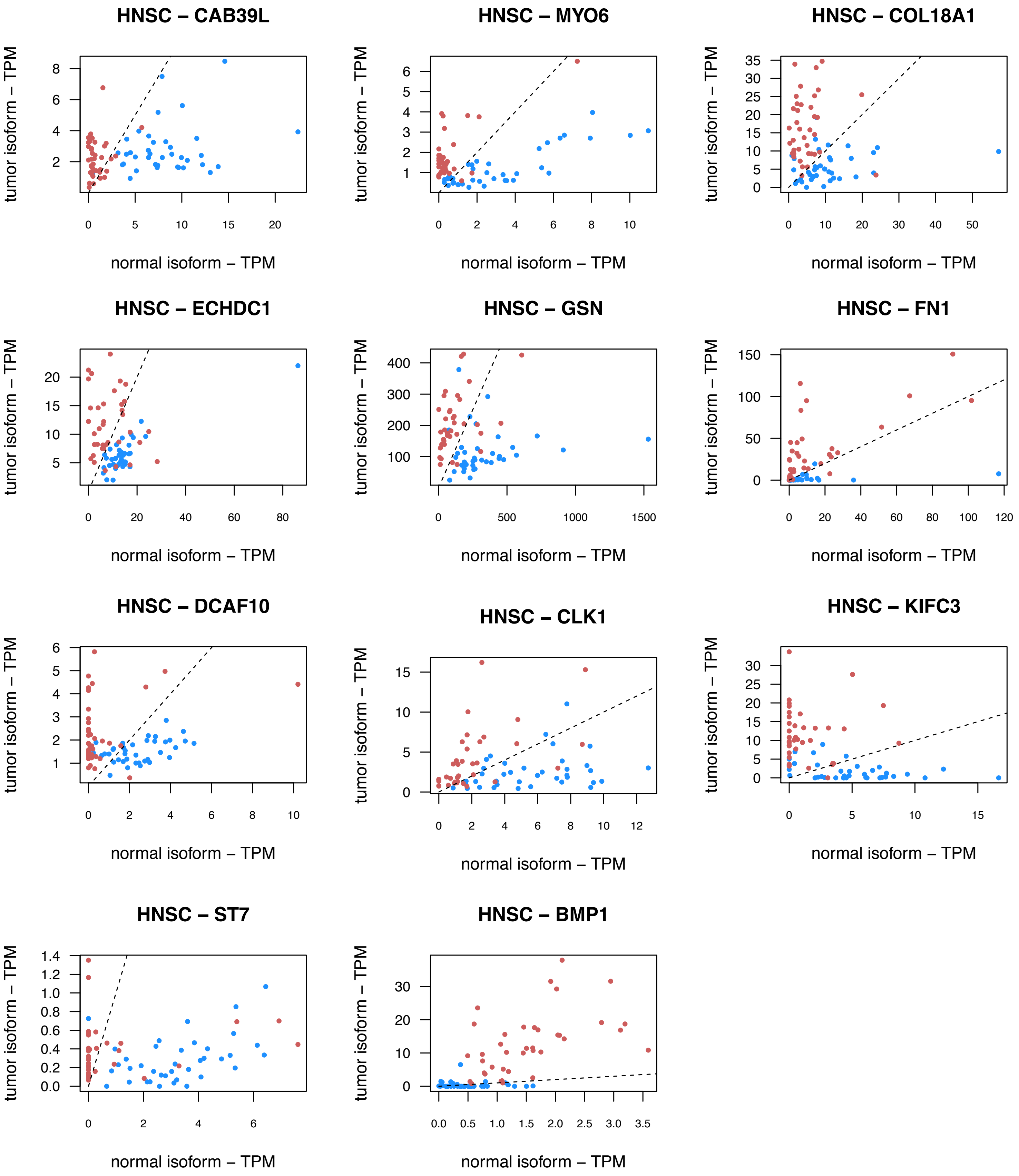


Supplementary Figure 2. **(A)** Isoform-pair models for HNSC, KICH, THCA and KIRC. Each panel shows the score S_1 and information gain (IG) for the selected models in each cancer type. All isoform pairs in the models are significant according to permutation analyses. Next to each cancer label, the expected average accuracy from the cross-validation analysis is given. **(B)** Test of the derived models on the set of held-out samples for the same cancer type. The tests were carried out on the unpaired tumor samples from (Table 1). Barplots indicate the proportion of samples (y-axis) for each number of possible correct votes (x-axis), i.e., the number of isoform-pair rules from the model. A sample is labelled according to majority vote, i.e. it is correctly if there more correct than incorrect isoform-pair rules.

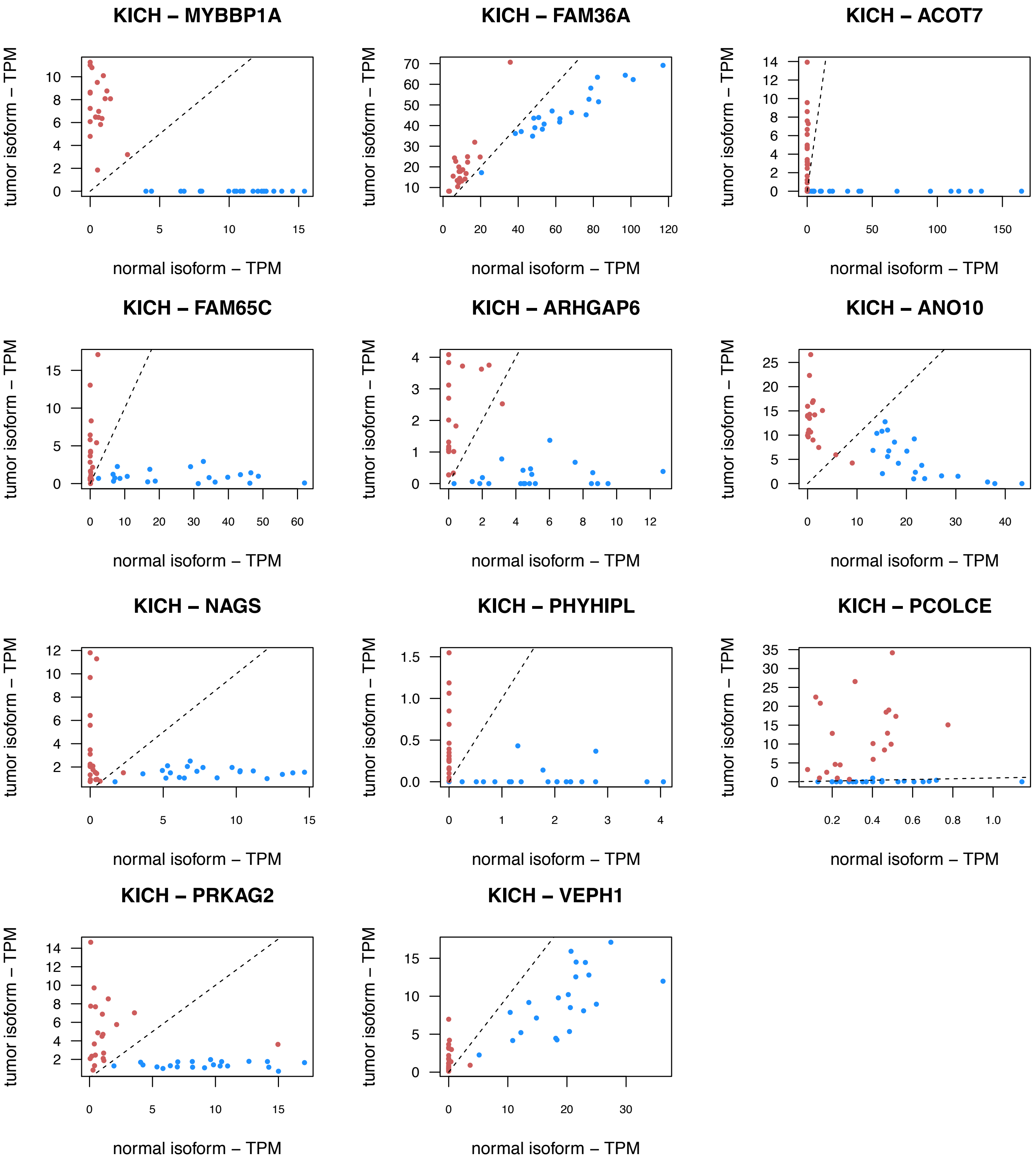
.



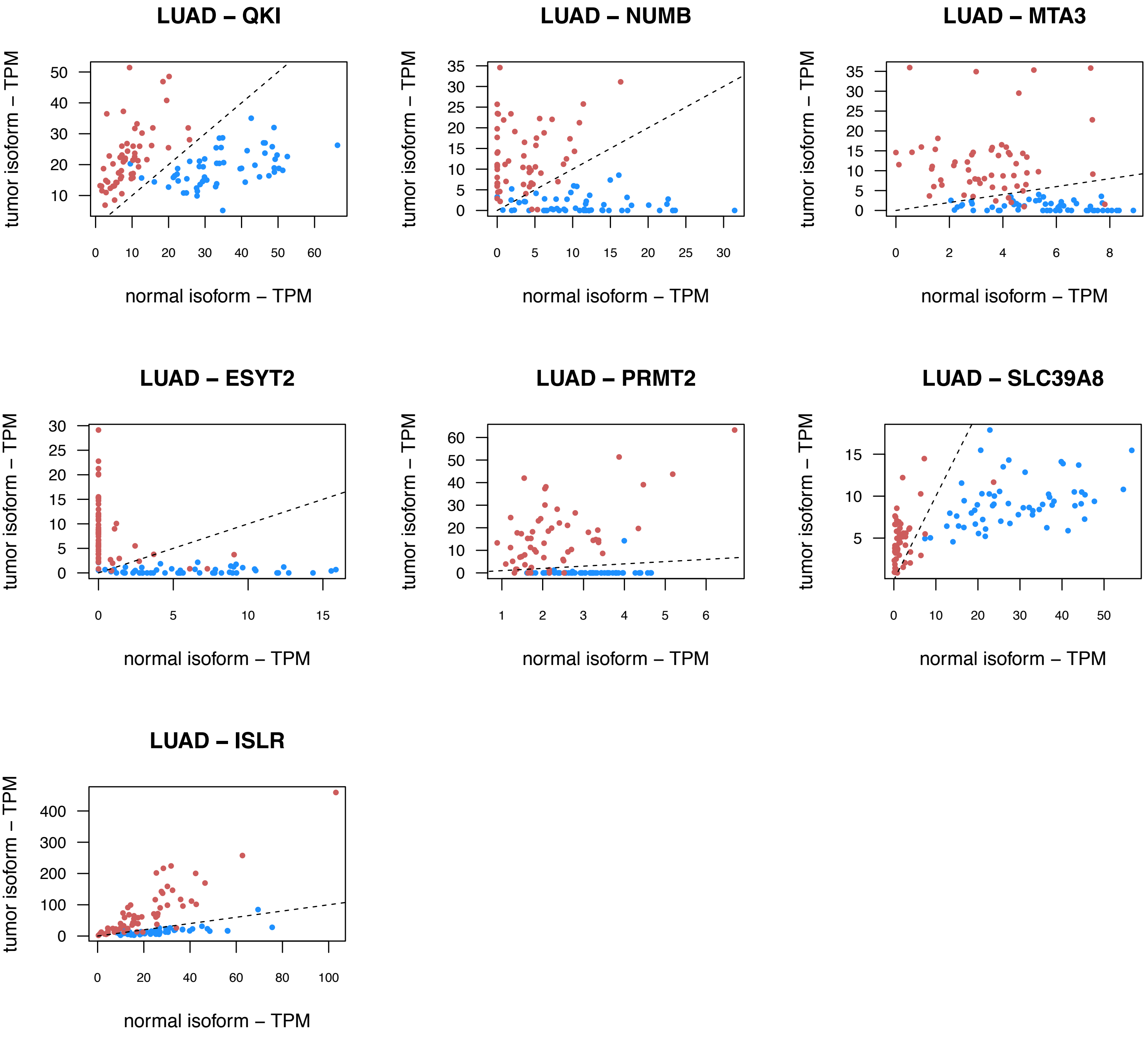
Supplementary Figure 3. Transcript expression plots, also called top scoring pair plots (Geman et al. 2004), for the isoform-pairs of the BRCA model **(A)** and the single-gene models of THCA, COAD and KIRC **(B)**. The plots show the expression for one isoform in the pair versus the expression for the second isoform in the pair for tumor (red) and normal (blue) samples. Expression values are given in transcript per million (TPM) units. The dashed line indicates the equal expression and represents the classification rule. The accuracy of the model is determined by how well the line separates the two samples.



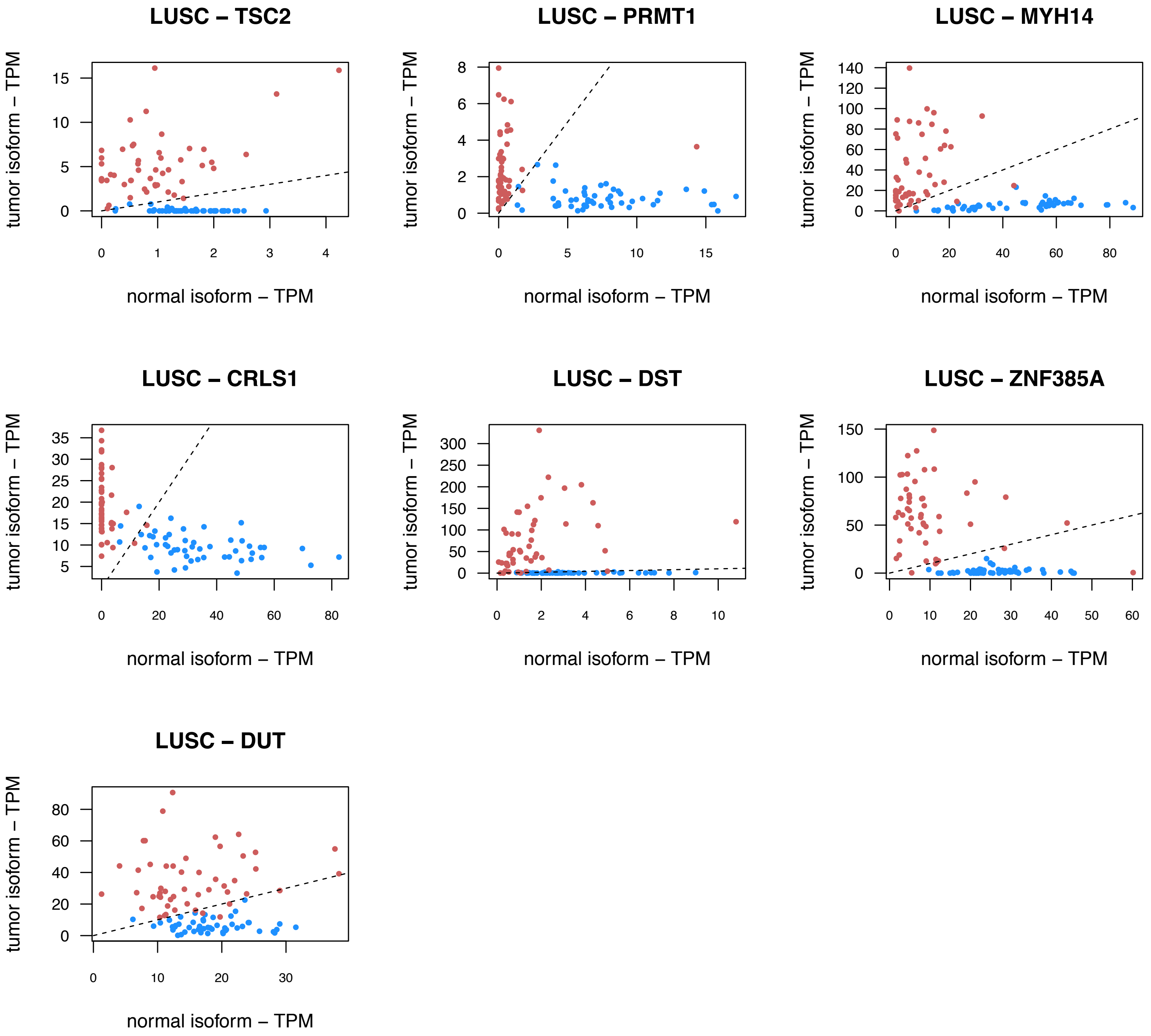
Supplementary Figure 4. Transcript expression plots for the isoform-pairs of the HNSC model The plots show the expression for one isoform in the pair versus the expression for the second isoform in the pair for tumor (red) and normal (blue) samples. Expression values are given in transcript per million (TPM) units. The dashed line indicates the equal expression and represents the classification rule. The accuracy of the model is determined by how well the line separates the two samples.



Supplementary Figure 5. Transcript expression plots for the isoform-pairs of the KICH model The plots show the expression for one isoform in the pair versus the expression for the second isoform in the pair for tumor (red) and normal (blue) samples. Expression values are given in transcript per million (TPM) units. The dashed line indicates the equal expression and represents the classification rule. The accuracy of the model is determined by how well the line separates the two samples.

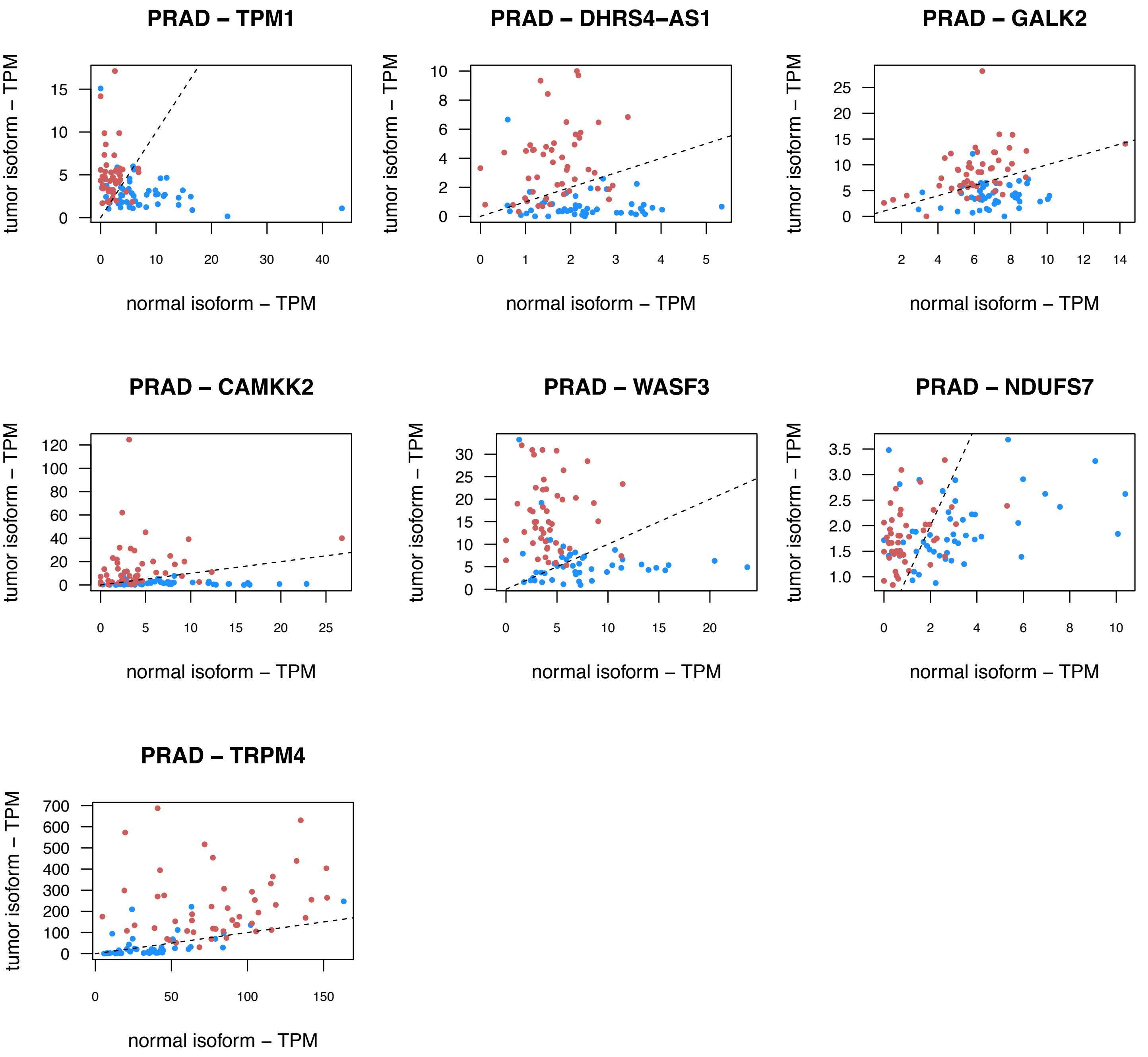


Supplementary Figure 6. Transcript expression plots for the isoform-pairs of the LUAD model. The plots show the expression for one isoform in the pair versus the expression for the second isoform in the pair for tumor (red) and normal (blue) samples. Expression values are given in transcript per million (TPM) units. The dashed line indicates the equal expression and represents the classification rule. The accuracy of the model is determined by how well the line separates the two samples.

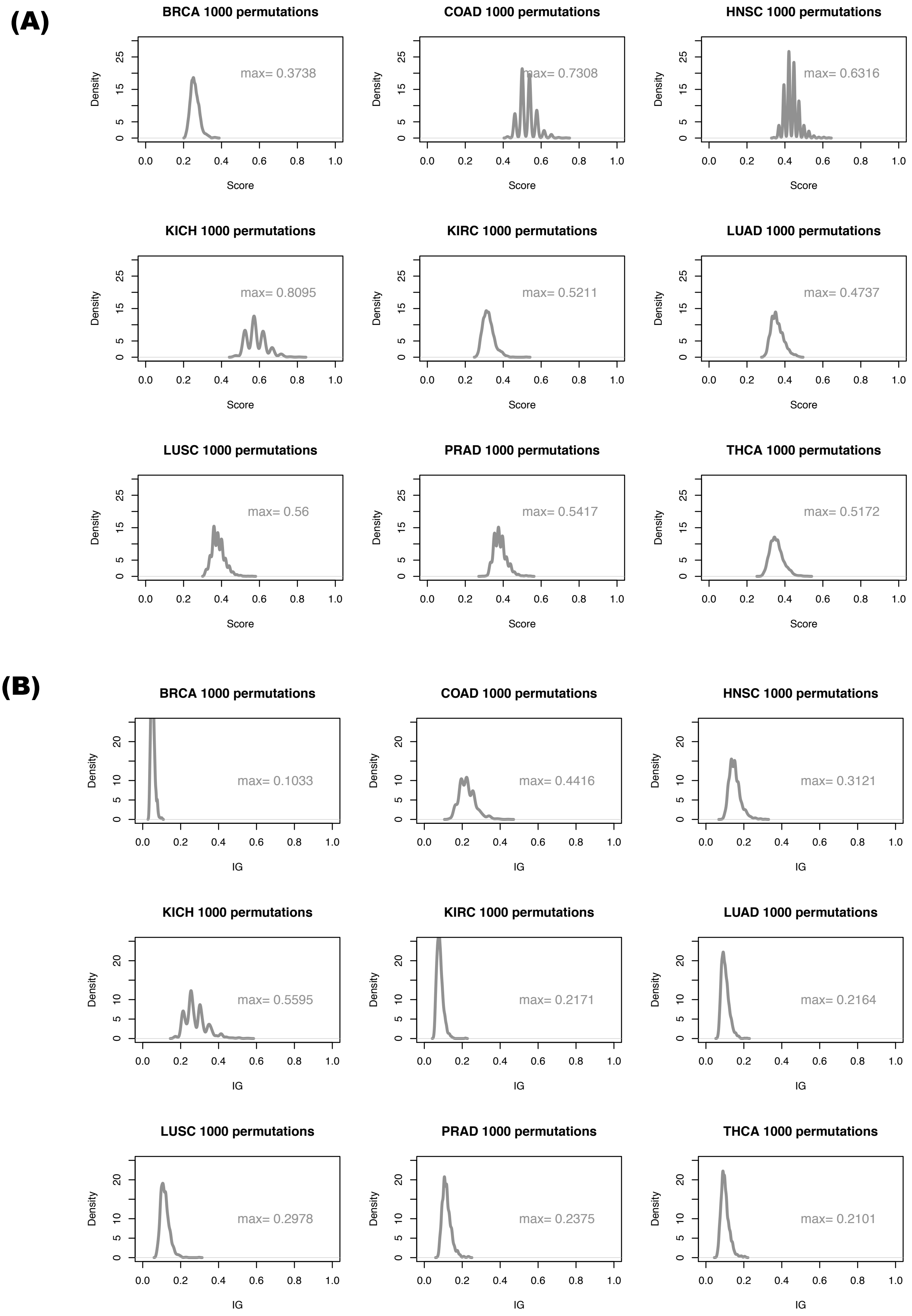


Supplementary Figure 7. Transcript expression plots for the isoform-pairs of the LUSC model. The plots show the expression for one isoform in the pair versus the expression for the second isoform in the pair for tumor (red) and normal (blue) samples. Expression values are given in transcript per million (TPM) units. The dashed line indicates the equal expression and represents the classification rule. The accuracy of the model is determined by how well the line separates the two samples.

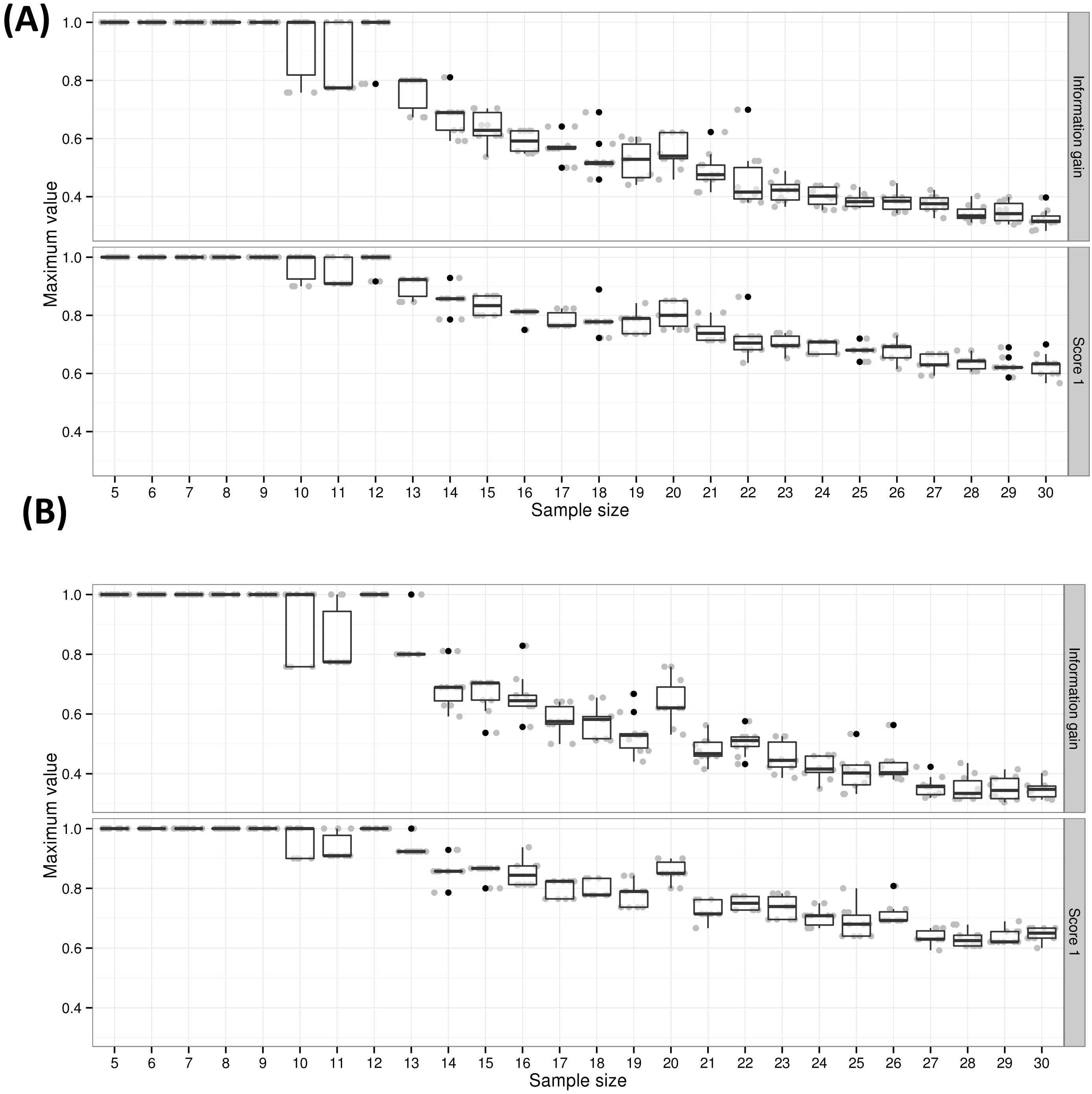
PRAD



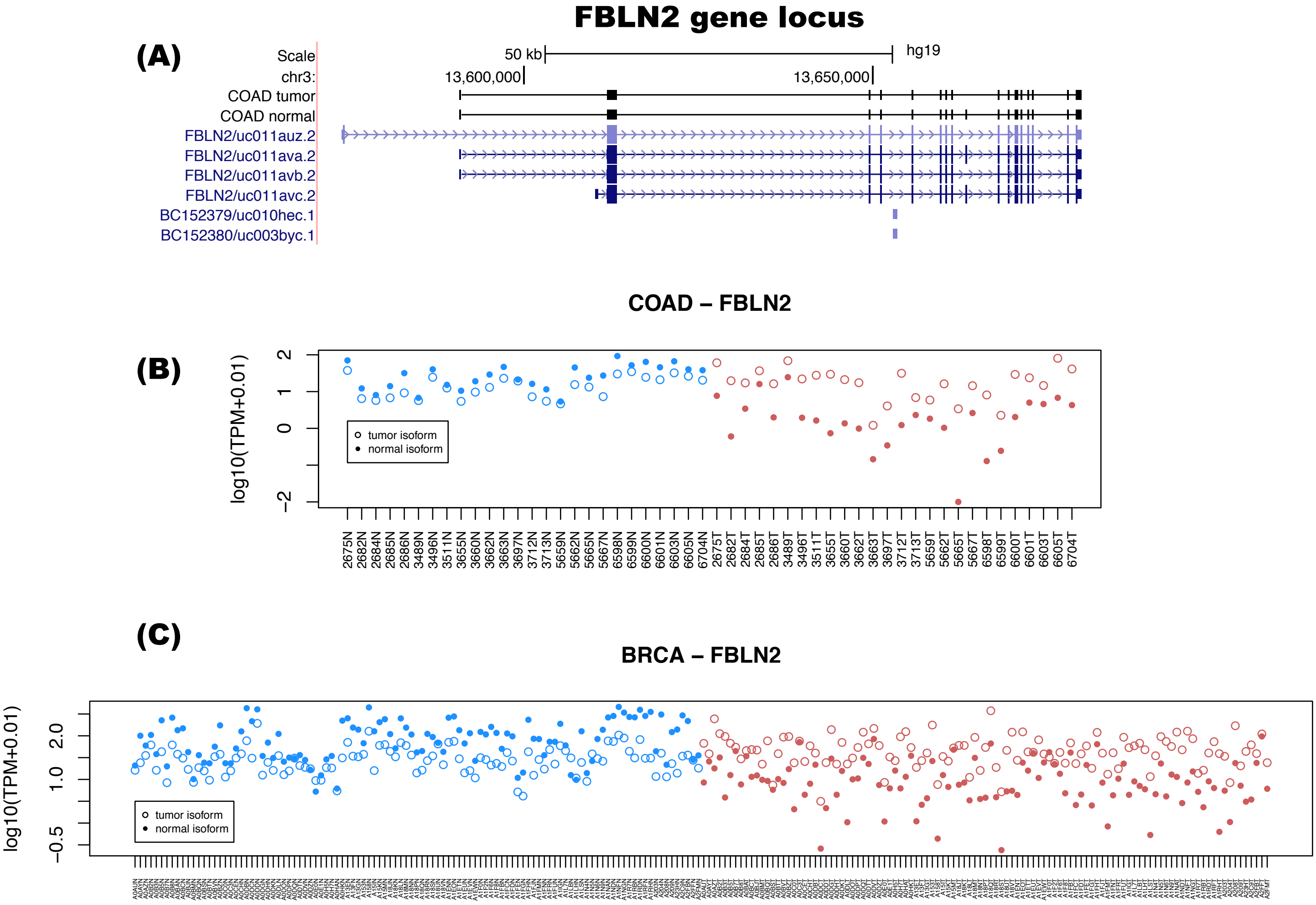
Supplementary Figure 8. Transcript expression plots for the isoform-pairs of the PRAD model. The plots show the expression for one isoform in the pair versus the expression for the second isoform in the pair for tumor (red) and normal (blue) samples. Expression values are given in transcript per million (TPM) units. The dashed line indicates the equal expression and represents the classification rule. The accuracy of the model is determined by how well the line separates the two samples. DHRS4-AS1 (C14orf167) is antisense of DHRS4



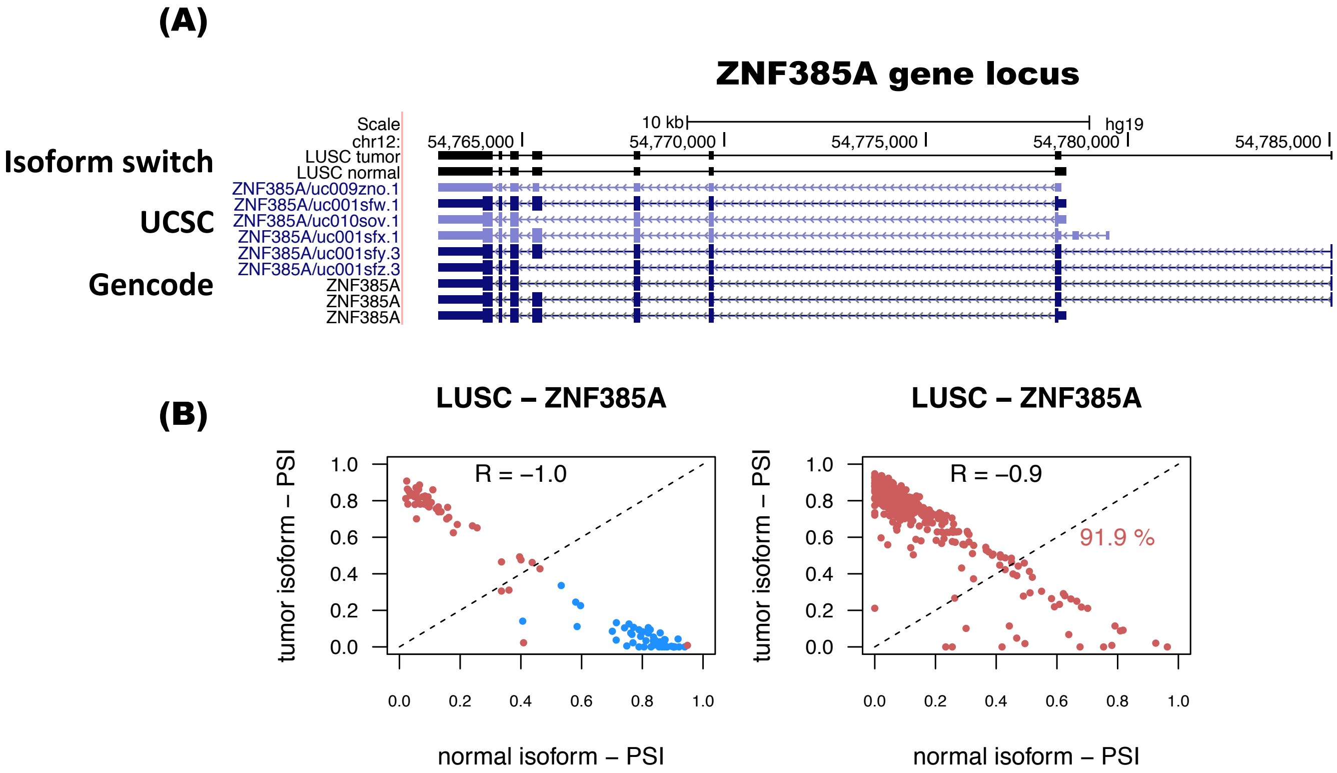
Supplementary Figure 9. Permutation analysis of the paired samples for the 9 cancer types. 1000 permutations of the labels were carried out, and at each permutation the Iso-kTSP algorithm was run and the top isoform-pair according to the cross-validation was selected. The plots show the distributions for the score S_1 (Score) **(A)** and Information Gain (IG) **(B)** for these 1000 top isoform-pairs of the permuted labels. An observed isoform-pair is considered significant if its individual score S_1 and IG values are larger than the maximum ones obtained from the 1000 permutations.



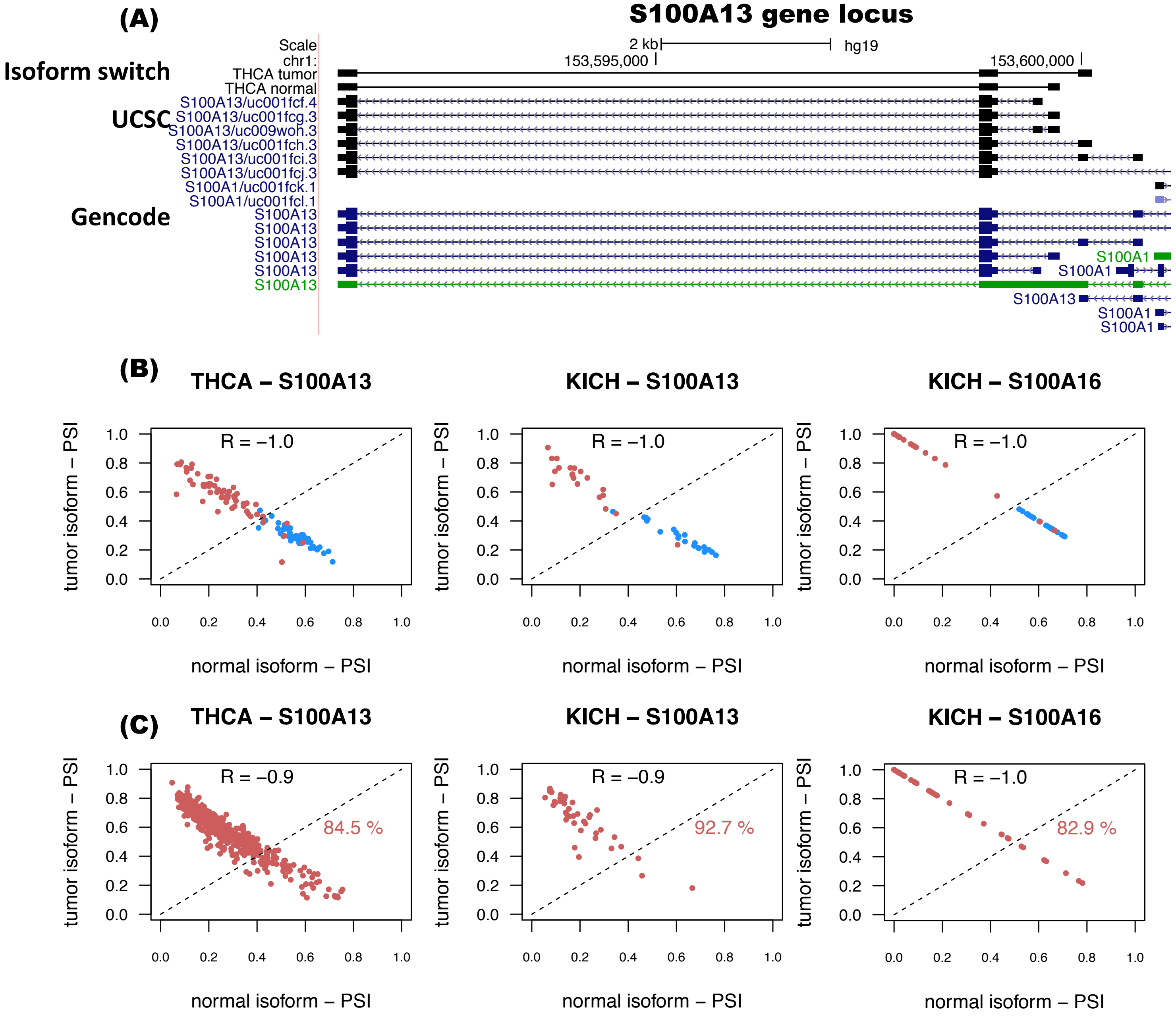
Supplementary Figure 10. Sample permutation with variable number of input samples. For BRCA **(A)** and LUSC **(B)**, we plot the distributions of the maximum values of the score S_1 and Information Gain for 10 permutations, using as input 5,6,… 30 paired input samples (x-axis). The plots indicate that below 12 paired samples (12 tumor and 12 normal), permuting the labels can produce with probability close to 1 isoform changes with score S_1 and IG close to 1. However, with 14 or more labels (13 for BRCA), permutation of the labels produces isoform changes by chance with smaller score 1 and IG values. Thus, significant isoform changes can be produced robustly with 14 or more paired samples.



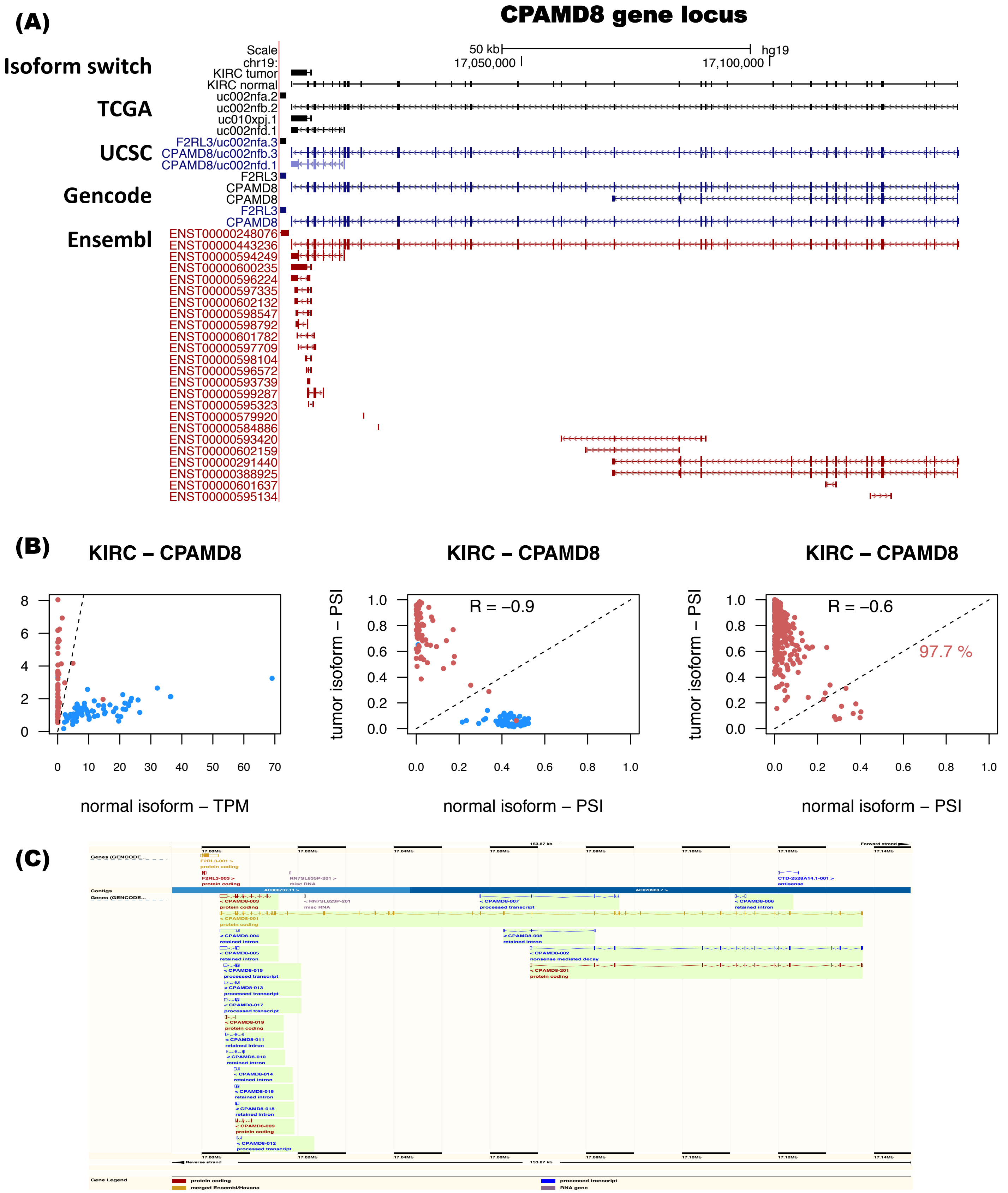
Supplementary Figure 11. (A) UCSC snapshot with the 2 isoforms involved in the isoform switch of FBLN2. The representation does not distinguish between protein coding or non-coding regions of the isoforms. The splicing change corresponds to a protein-coding cassette exon. Representation of the TPMs in log10 scale (y-axis) for both FBLN2 isoforms in COAD **(B)** and BRCA **(C)** samples (x axis)



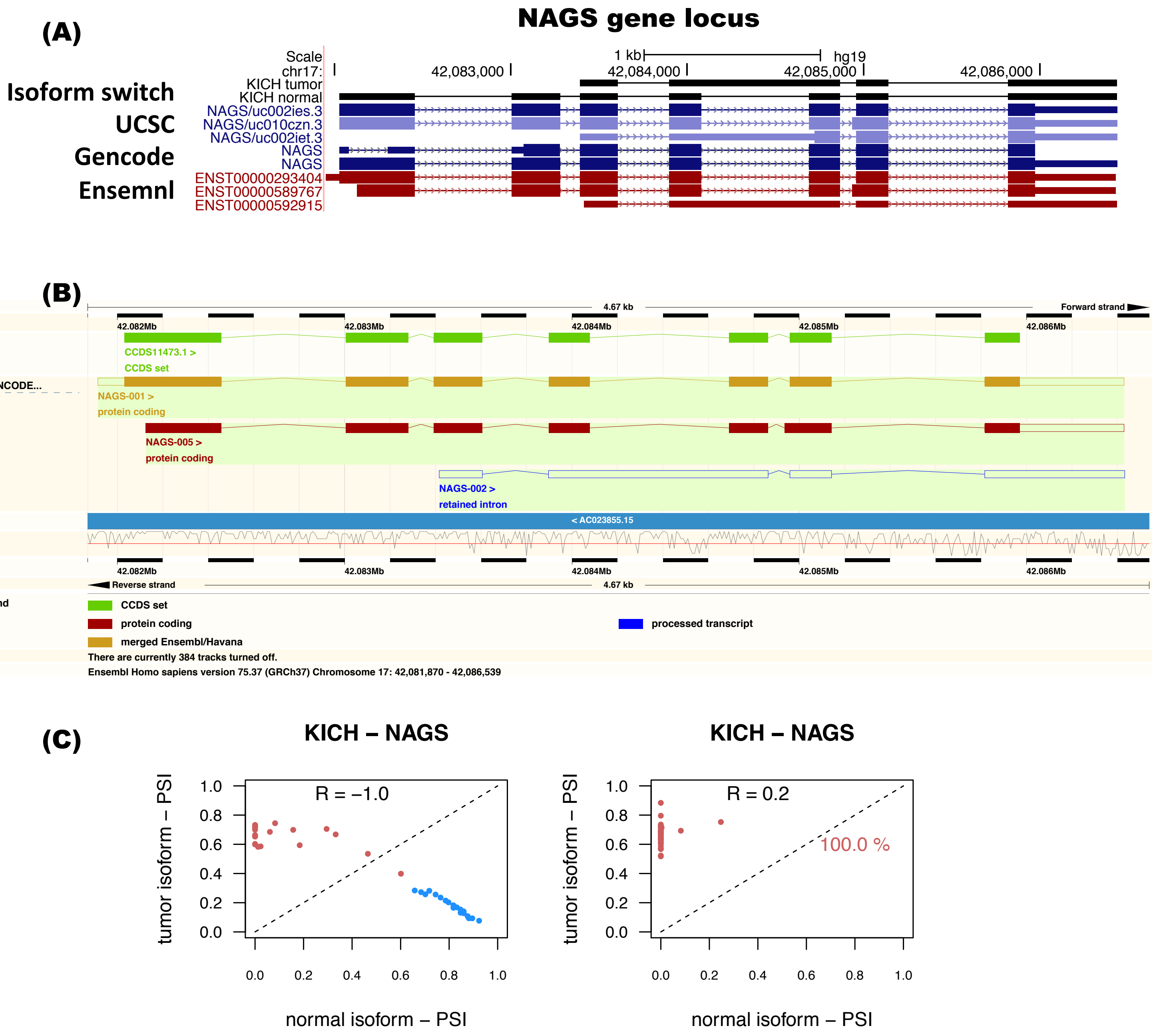
Supplementary Figure 12. There is an isoform switch in LUSC in the gene ZNF385A, which is related to the use of alternative first exon and alternative 3' splice-site. This gene produces protein interacts with TP53 enhancing its function in cell-cycle arrest, thereby promoting growth arrest (Das et al. 2007). Our representation of the isoform-pair does not distinguish between protein coding or non-coding regions of the isoforms **(B)** Plots the PSI values for the ZNF385A isoform-pairs (left panel) and unpaired tumor samples (right panel) in LUSC. Values in tumor samples are in red, whereas values in normal samples are in blue. The percentage of correctly classified unpaired tumor samples is given.



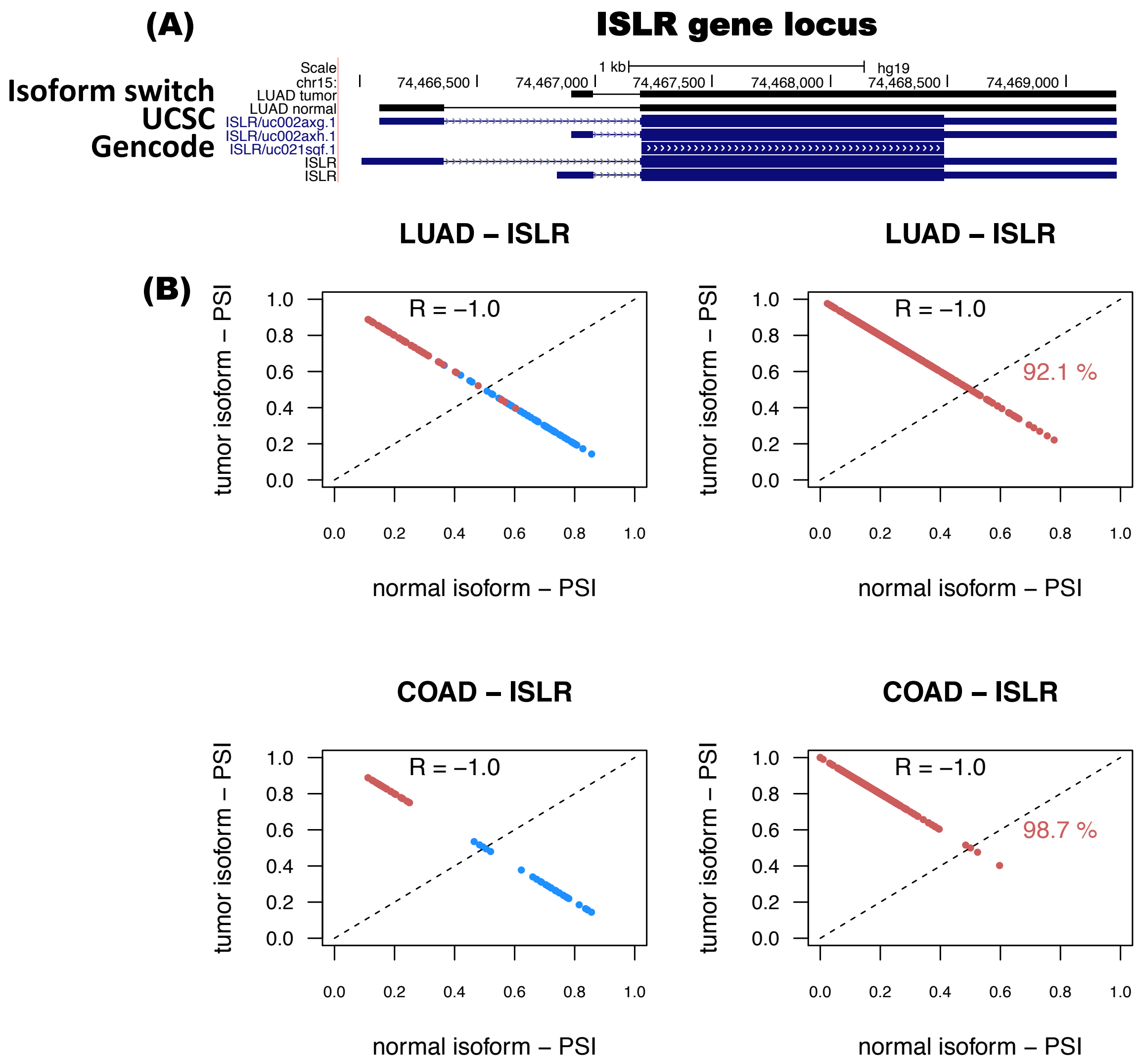
Supplementary Figure 13. (A) The gene S100A13, coding for a calcium binding protein, has multiple isoforms, all differing in the 5’ UTR region. The predicted isoform change in S100A13 involves two anti-correlating isoforms differing in a single alternative first 5’UTR exon. S100A13 has been proposed to be a new marker of angiogenesis in various cancer types (Massi et al. 2010). Interestingly, S100A13 and another member of the S100 family, S100A16, have an isoform switch in KICH **(B)** Plots for the relative inclusion of the isoform pairs for S100A13 and S100A16 in unpaired samples The two isoforms show anticorrelation in their relative inclusion levels. **(C)** PSI anticorrelation in the unpaired samples. The S100A13 isoform change alone classifies correctly 84.5% and 92.7% of the THCA and KICH unpaired tumor samples, respectively. Similarly, S100A16 can separate correctly 82.9% of the KICH unpaired samples.



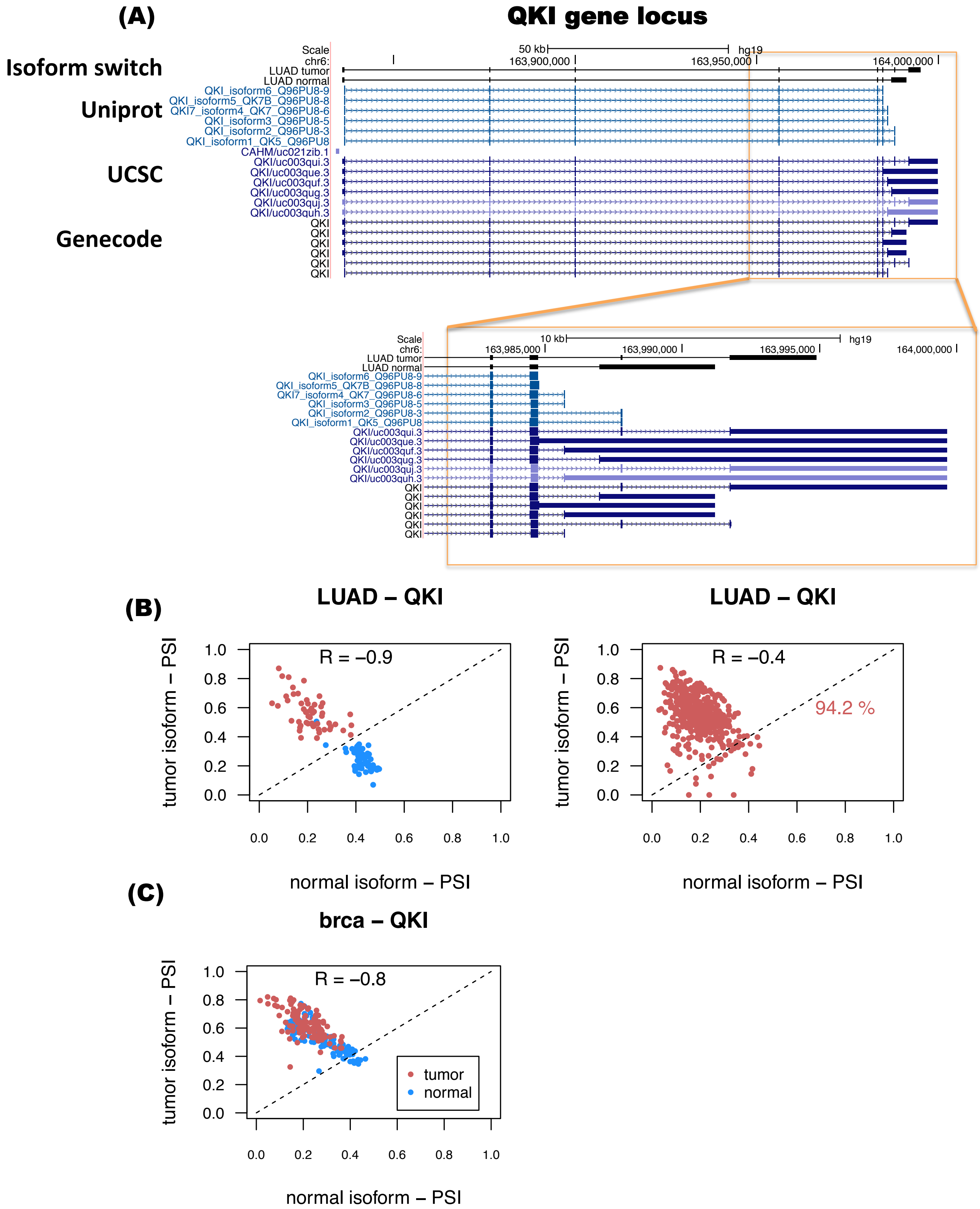
Supplementary Figure 14. CPAMD8 locus. (A) KIRC model is composed of a single isoform switch in CPAMD8, which is implicated in innate immunity and damage control (Li et al. 2004). **(B)** The two isoforms involved show anti-correlation and **(C)** the predicted change gives rise to a truncated isoform that contains a retained intron, however the normal isoform does not reach PSI = 1.



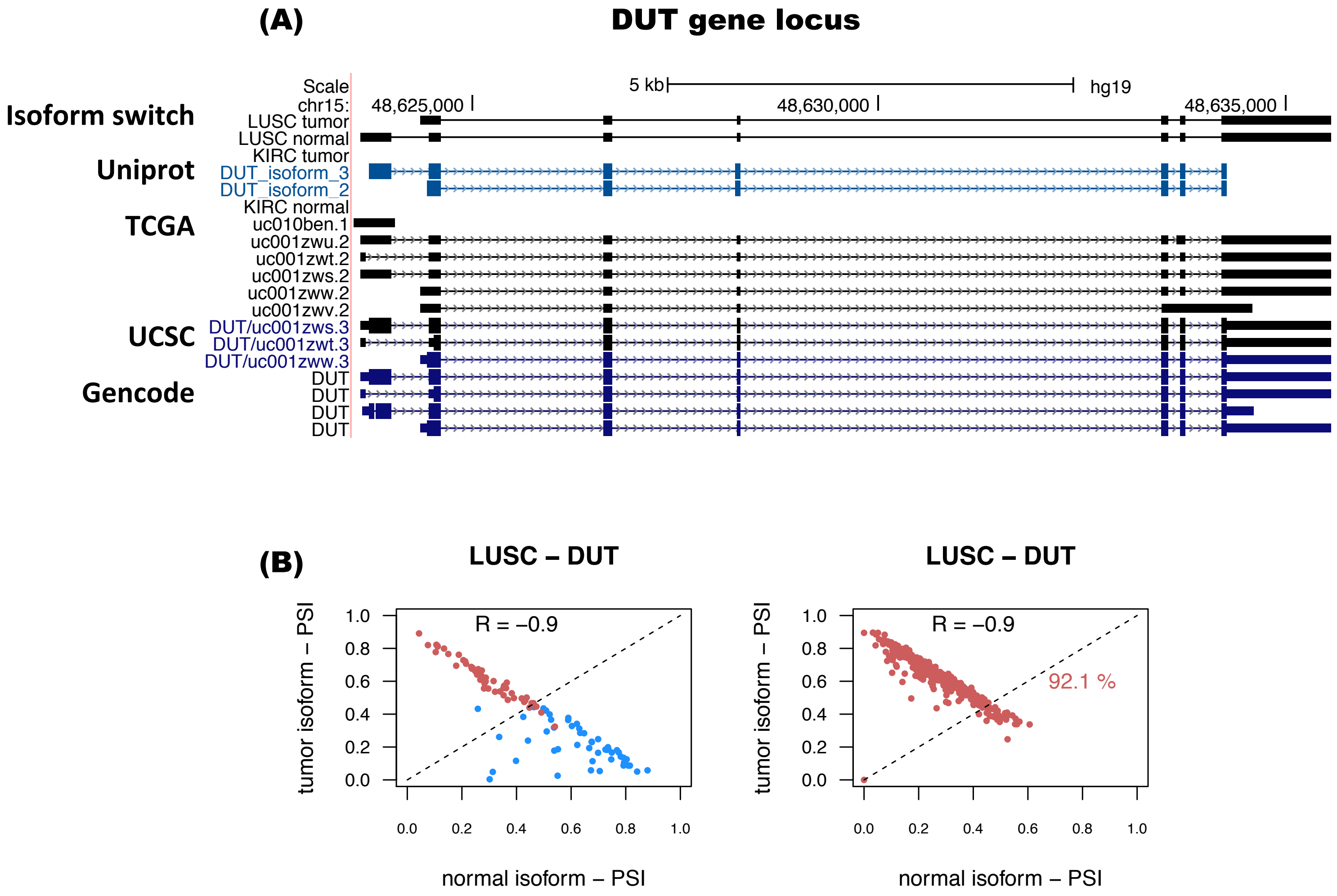
Supplementary Figure 15 .(A) The isoform switch in the N-acetylglutamate synthase involves an intron retention between the 4th and 5th protein exons in KICH tumor samples, which changes a coding region into non-coding in the tumor isoform. **(B)** This transcript is annotatedby Ensembl and Havana as an intron-retention without protein product (NAGS-002 ENST00000592915). **(C)** The relative inclusion of the two isofroms anticorrelate in paired and unpaired samples (left and right panels, respectively). In the unpaired samples, 100% of the KICH tumor samples agree with the found isoform rule and show that the normal isoform is not used in most of the samples.



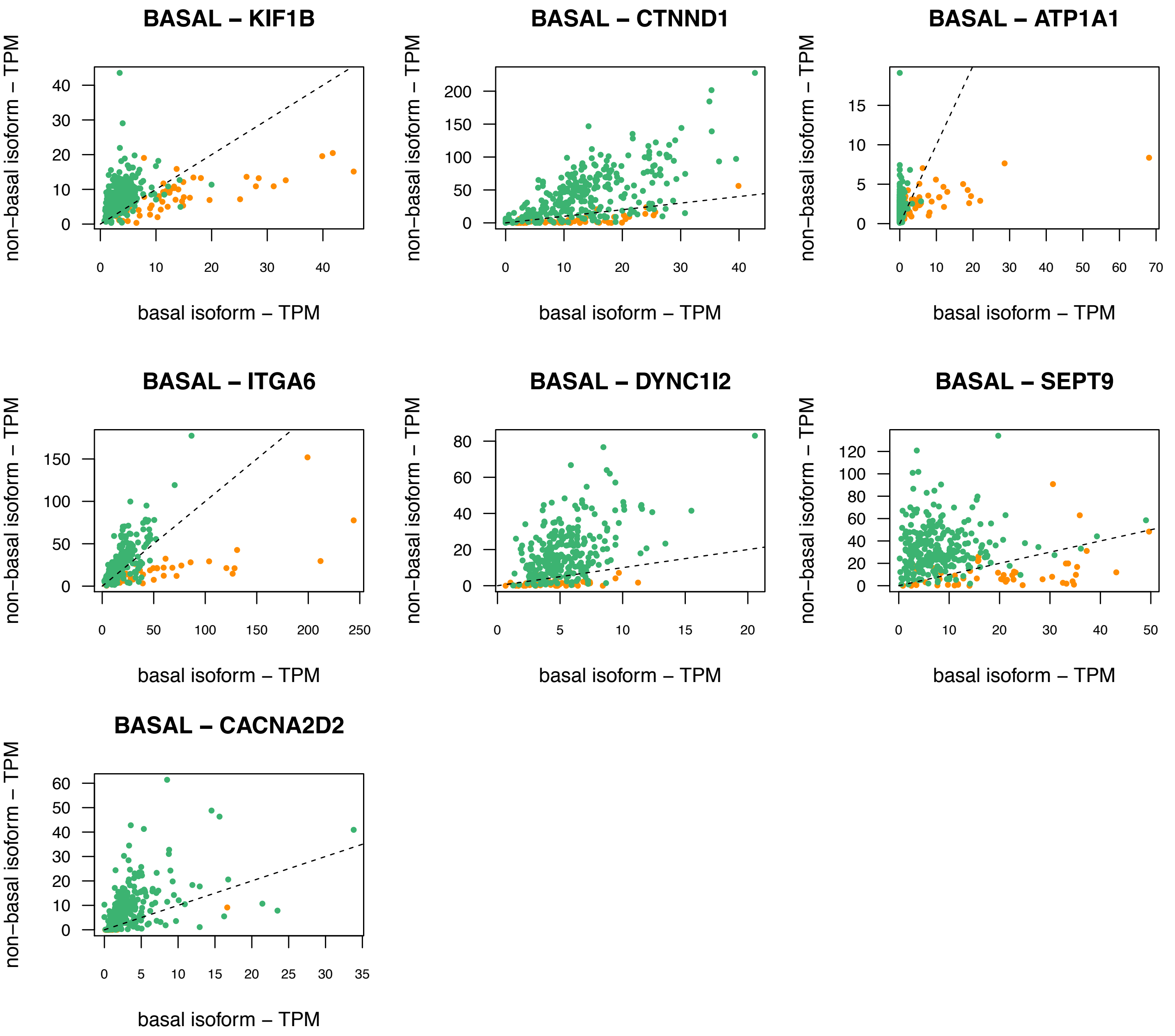
Supplementary Figure 16: LUAD and COAD include a switch in the gene ISLR, involving a change in the first exon **(A)** which shows a perfect anticorrelation in paired and unpaired samples **(B)**. In this case, the strong anticorrelation is due to this gene having only two isoforms. To the best of our knowledge, the activity of this gene has not been related to cancer so far.



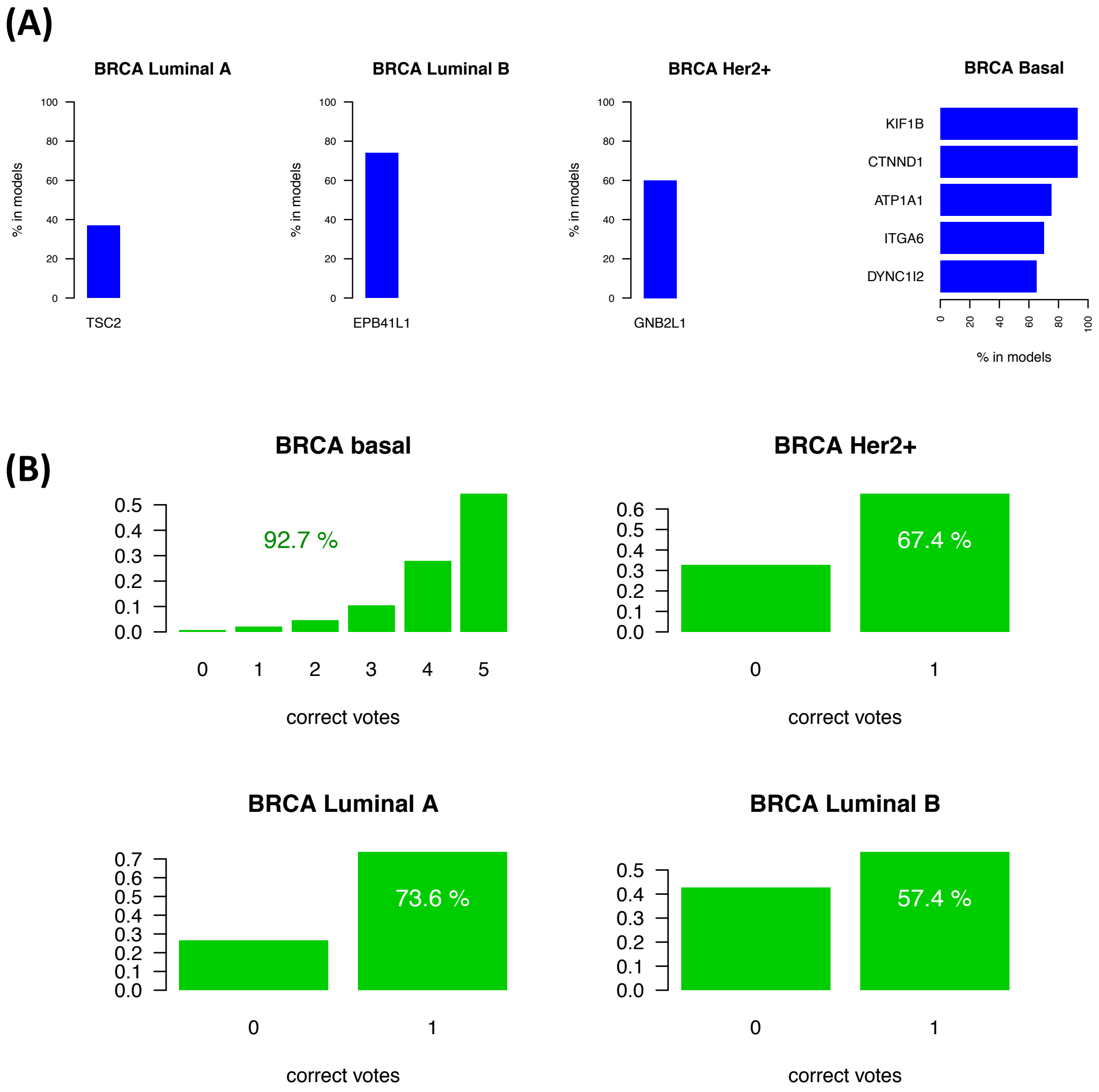
Supplementary Figure 17. (A) The tumoral isoform (uc003qui.2) coincides with NM_006775, which codes for the protein Q96PU8, often denoted as QKI-5. The normal isoform (uc003qug.2) coincides with RefSeq NM_206853, which encodes for protein Q96PU8-9.**(B)** The two isoforms show an anticorrelation pattern. **(C)** The same two isoforms show anticorrelation in BRCA samples, but they show similar inclusion.



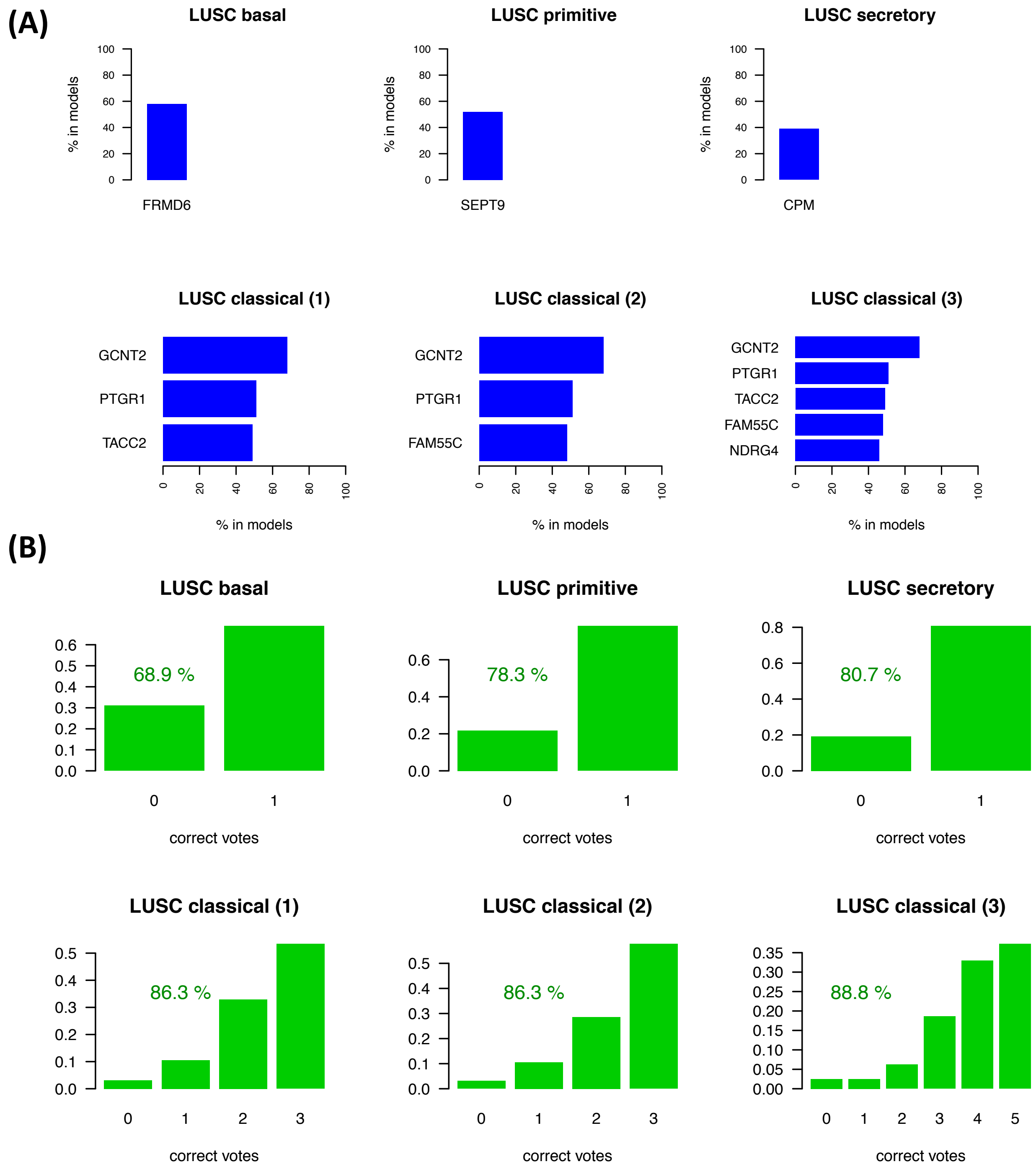
Supplmentary Figure 18. (A) DUT gene has two isoforms, named isoform 2 (Uniprot identifier P33316-2) and isoform 3 (Uniprot identifier P33316). The track "Uniprot" contains the mapping of these two proteins to the DUT genome locus. The isoform switch predicts the RNA encoding for protein isoform 2 to be more included in tumor samples. Protein Isoform 2 has been involved in the onset of DNA replication in lung fibroblasts (Ladner et al. 1997). **(B)** The two isoforms show anti-correlation in their PSIs in paired normal samples (left) and in unpaired samples (right).



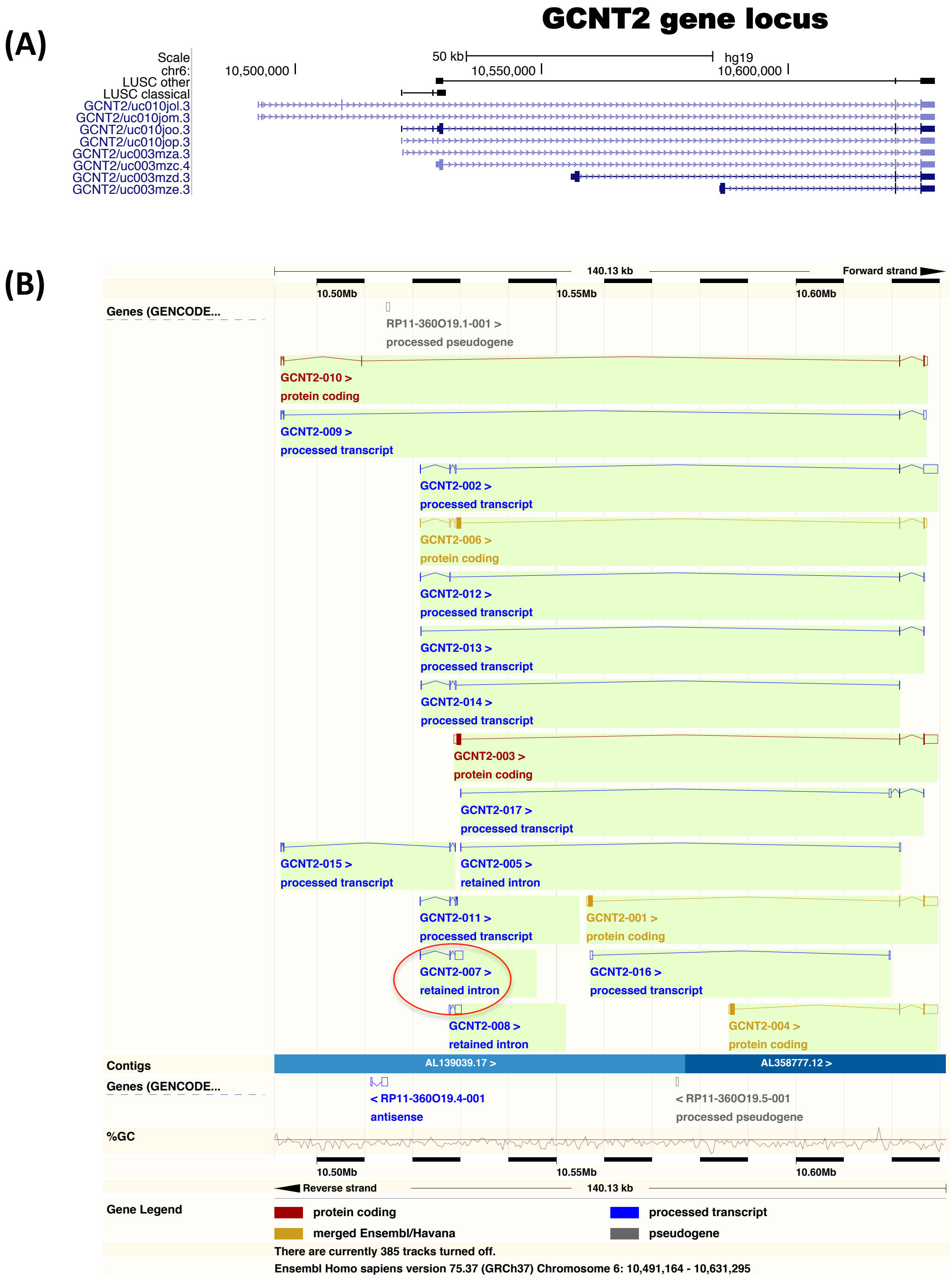
Supplementary Figure 19. Transcript expression plots for the Isoform-pair model for basal-like BRCA tumors. The plots show the TPM for the isoform pairs found to be specific to BRCA basal samples compared to the other BRCA subtypes (luminal A, luminal B and Her2+). We show those that show significance in more than 80% of the balanced comparisons using random subsampling. Basal samples are depicted in orange, whereas the samples from the other subtypes are shown in green.



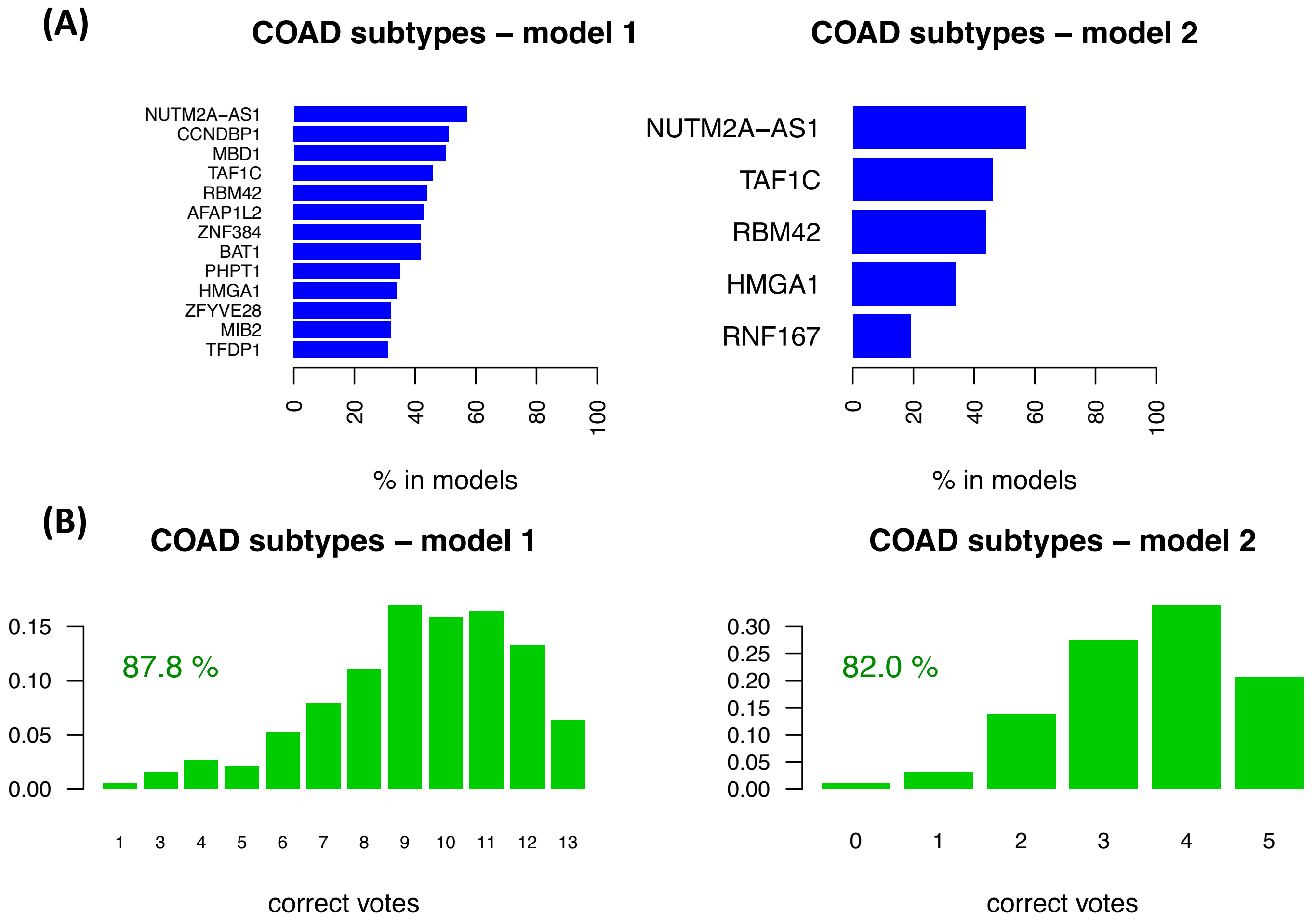
Supplementary Figure 20. Calculated isoform change models for BRCA subtypes (A) Models obtained for the four BRCA subtypes: luminal A, luminal B, Her2+ and basal. The basal model shows the top 5 isoform pairs. **(B)** Performance each model on the entire dataset. The barplots show the proportion of tested samples with a specific number of votes correct. A true prediction is considered when the majority of votes are correct. The printed percentage refers to the total number of samples that are correctly classified. Basal shows an overall accuracy of 92.% with more than 50% of the samples agreeing with the 5 isoform-pair rules.



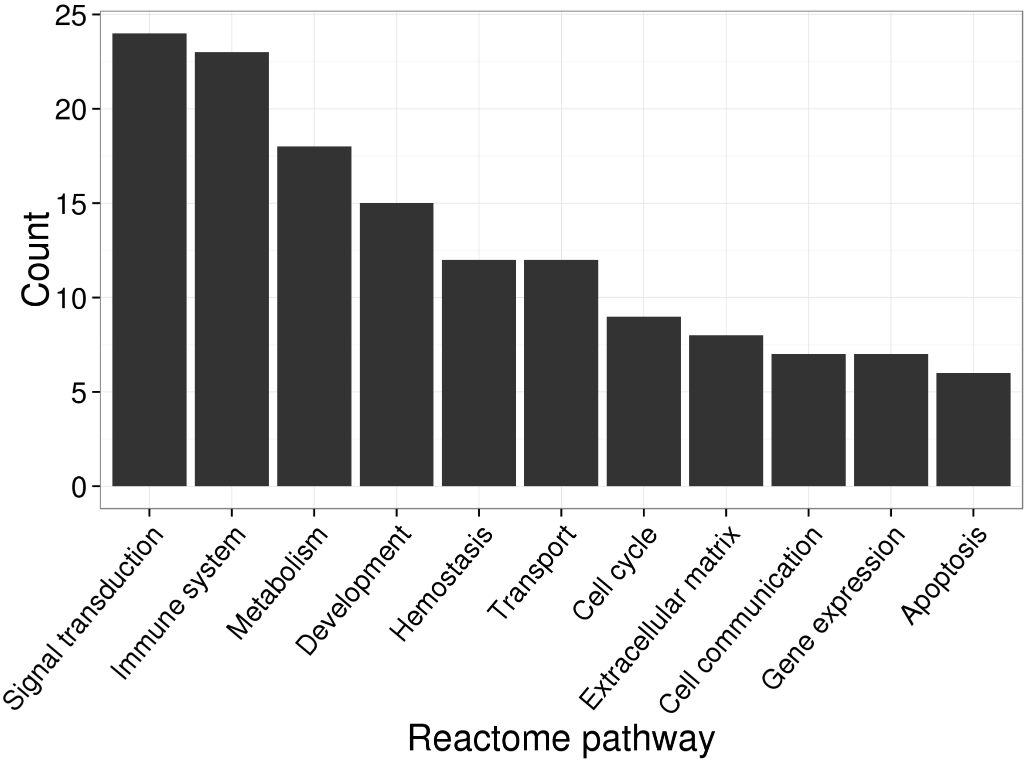
Supplementary Figure 21. Calculated isoform change models for LUSC subtypes. (A) Basal, primitive and secretory show almost no recurrent changes. We considered the top ones for testing. . For the classical subtype we considered 3 models with 3 and 5 isoform changes that include those genes that appeared in more than 40% of the models: PTGR1, TACC2, FAM55C and NDRG4. Except for GCNT2, which was significant in 22% of the iterations, all other isoform changes were significant in at most 3% of them. The figure indicate the fraction of random subsampling iterations (comparing one subtype against a balanced pool from the rest) for which each gene was selected in a predictive model. **(B)** Performance of the models when tested on the entire LUSC tumor set, comparing each subtype vs all the rest. The barplots show the number of samples with each possible number of correct votes (from 0 to the number of genes in the model), and the percentage of samples correctly classified.



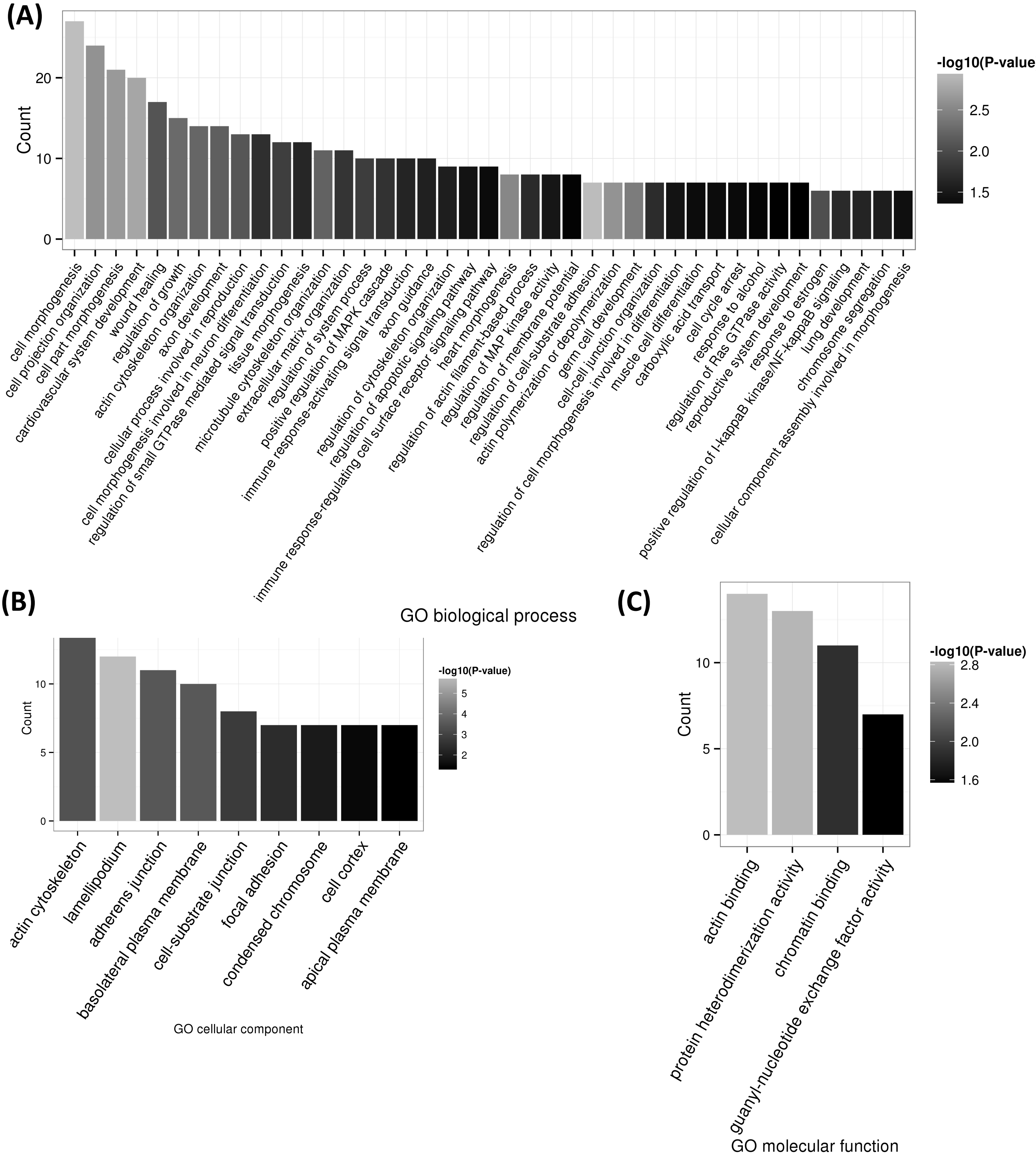
Supplementary Figure 22. (A) Isoform change predicted for GCNT2 between the classical and the other LUSC subtypes. Isoforms are represented in black, without indicating coding or non-coding regions. **(B)** The isoform frequently found in the classical samples correspond to an isoform annotated as non-coding and with a retained intron by Ensembl.



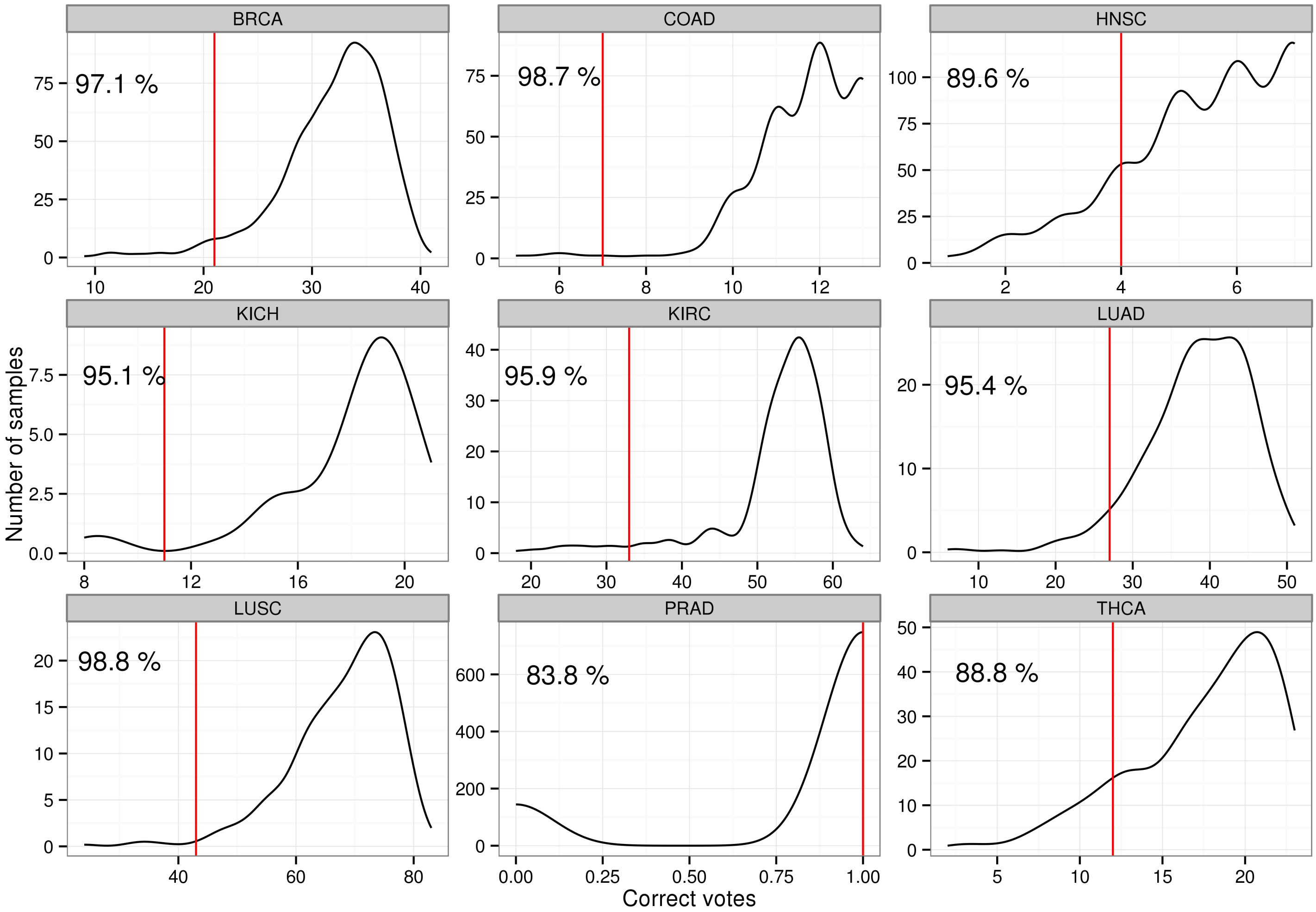
Supplementary Figure 23. Calculated isoform changes between COAD hyper and non-hyper mutated samples (A) We considered two models with 13 and 5 isoform-pairs that appeared in models for more than 20% of the iterations. None of the found isoform-pairs were significant for more than 4% of the samples. **(B)** Performance of the two models when tested on the entire COAD tumor set, separating hyper and non-hyper mutated samples. The barplots show the number of samples with each possible number of correct votes (from 0 to the number of genes in the model), and the percentage of samples correctly classified.



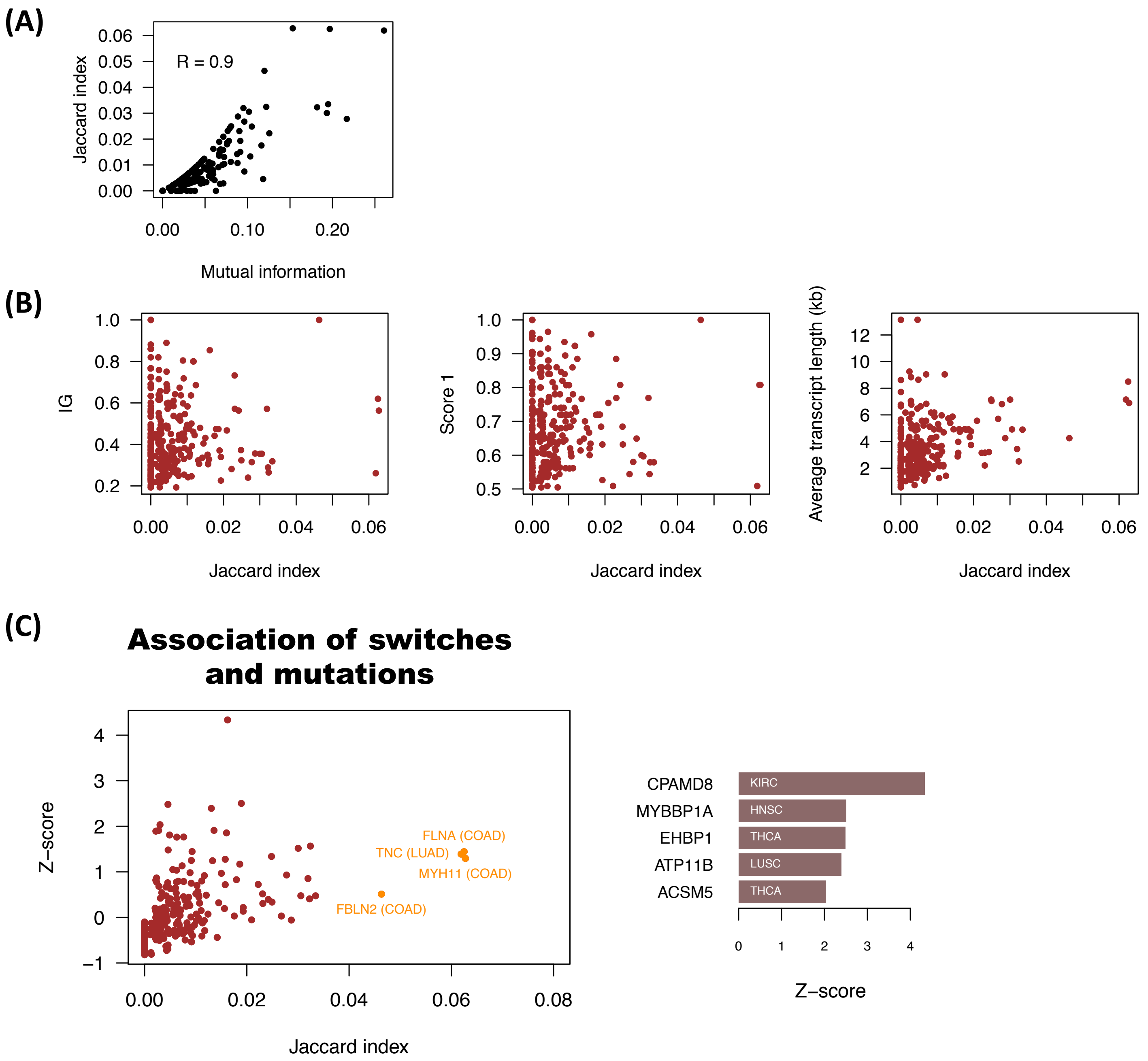
Supplementary Figure 24. Reactome pathway analysis. The barplot shows the most frequent Reactome pathways represented by the 244 isoform switches (Figure 5), based on the v68 of the Reactome database. None of these pathways show a significant enrichment compared to the total number of genes with multiple isoforms as background.



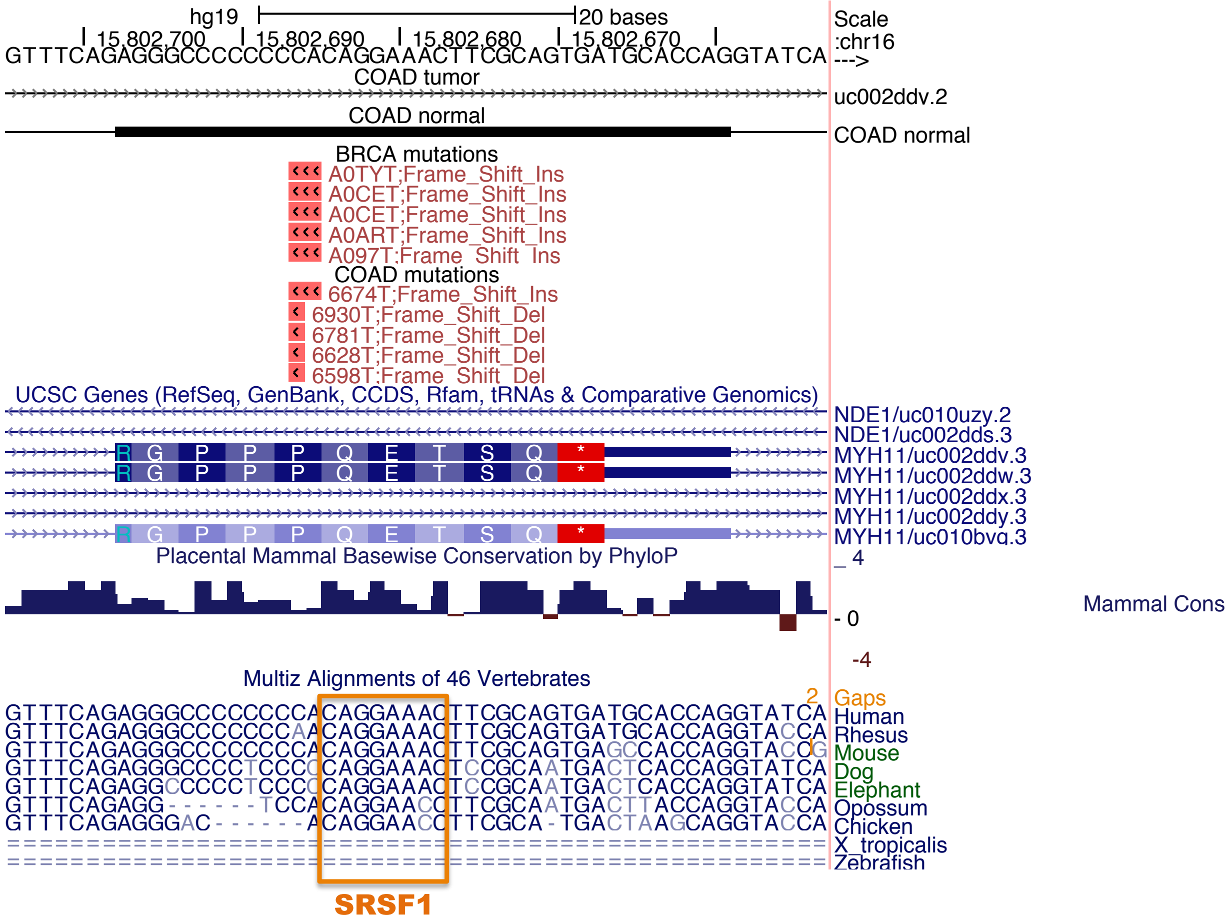
Supplementary Figure 25. GO enrichment analysis. The barplots show the most frequent GO terms represented by the 244 isoform switches for Biological Process **(A)**, Cellular Component **(B)** and Molecular Function **(C)** ontologies. The GOstat package was used for the test, with a P-value cutoff of 0.05, odds-ratio > 2 and at least 5 genes. The background set was the gene set used for the Iso-kTSP analysis and the "conditional" option was used in the test to calculate the significance of a parent term, if its children were significant. All the shown categories are significant. The gray scale indicates the –log10(p-value).



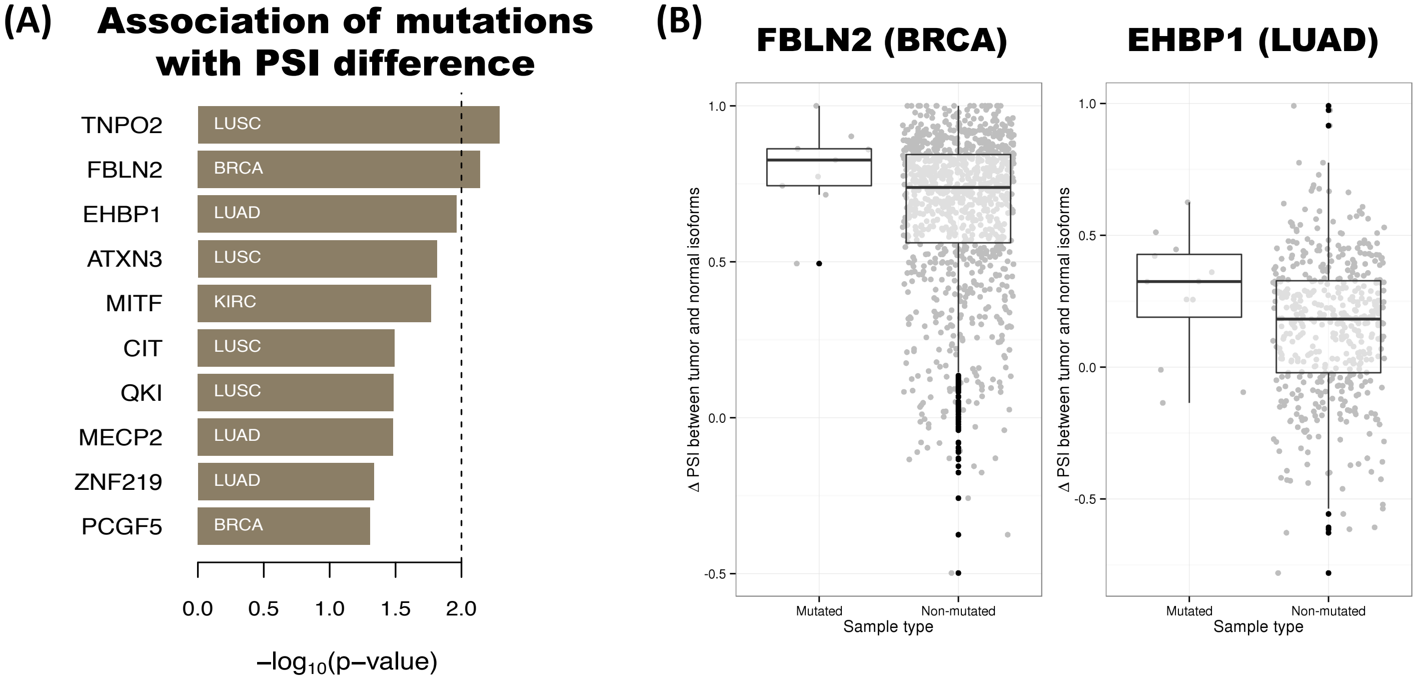
Supplementary Figure 26. Performance of the isoform switches on the unpaired tumor samples. For each cancer type, we tested the models built by using all significant isoform switches found in that cancer type. In each we considered all switches if they were an odd number, or removed the last one from the ranking of Information Gain if they were an even number. The density plots show the number of samples (y axis) according to the number of possible correct votes (x-axis). The vertical red line indicates the number of correct votes necessary for correct labeling of a sample (= (|switches|+1)/2). PRAD is a special case as it only has one isoform switch. The percentage of samples correctly classified is also indicated.



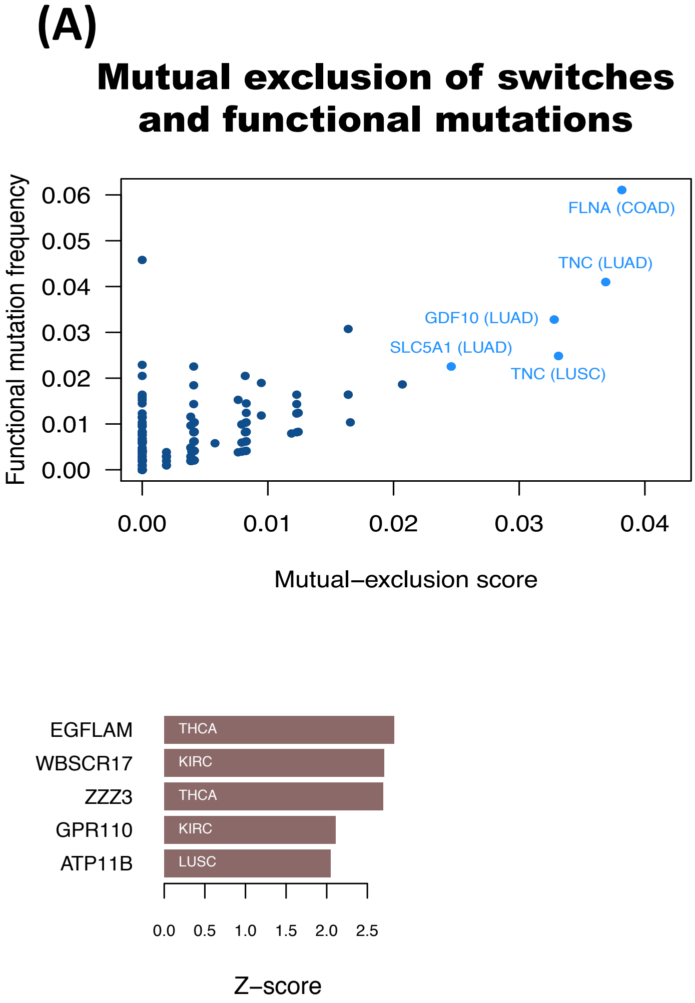
Supplementary Figure 27. (A) Correlation between Jaccard scores and mututal information. In both cases the association is calculated across samples, using per sample the presence or absence of the switch and the presence or not of one or more mutations. **(B)** Relation between the Jaccard index and three important variables of the isoform switches: Information gain of the switch, score S_1 and transcript length (averaged for both isoforms in the pair). **(C)** Relation between Jaccard indexes and the corresponding z-scores, calculated by comparing to the Jaccard index in the same cancer type for a set of 100 random transcripts with similar total exonic length. Z-scores (right barplot) tend to be higher for switches in cancer with very few mutations. That is, the top z-scores are due to 1 or 2 mutated samples only. These plots were performed with all mutation types falling in isoforms. Similar results are obtained when using all mutations falling on the full gene region, and when using only protein-affecting mutations on isoforms or gene-regions.


hg19
20 bases
Scale
15,802,700
15,802,690
15,802,680
15,802,670
chr16
GTTTCAGAGGGCCCCCCCACAGGAAACTTCGCAGTGATGCACCAGGTATCA
--->
COAD tumor
uc002ddv.2
COAD normal
COAD normal
BRCA mutations
A0TYT;Frame_Shift_Ins
A0CET;Frame_Shift_Ins
A0CET;Frame_Shift_Ins
A0ART;Frame_Shift_Ins
A097T;Frame_Shift_Ins
COAD mutations
6674T;Frame_Shift_Ins
6930T;Frame_Shift_Del
6781T;Frame_Shift_Del
6628T;Frame_Shift_Del
6598T;Frame_Shift_Del
UCSC Genes (RefSeq, GenBank, CCDS, Rfam, tRNAs & Comparative Genomics)
NDE1/uc010uzy.2
NDE1/uc002dds.3
R G P P P Q E T S Q *
MYH11/uc002ddv.3
R G P P P Q E T S Q *
MYH11/uc002ddw.3
MYH11/uc002ddx.3
MYH11/uc002ddy.3
R G P P P Q E T S Q *
MYH11/uc010bvq.3
Placental Mammal Basewise Conservation by PhyloP
4
Mammal Cons
0
-4
Multiz Alignments of 46 Vertebrates
2
Gaps
GTTTCAGAGGGCCCCCCCACAGGAAACTTCGCAGTGATGCACCAGGTATCA Human
GTTTCAGAGGGCCCCCCAACAGGAAACTTCGCAGTGATGCACCAGGTACCA Rhesus
GTTTCAGAGGGCCCCCCCACAGGAAACTTCGCAGTGAGCCACCAGGTACCG Mouse
GTTTCAGAGGGCCCCTCCCCAGGAAACTCCGCAATGACTCACCAGGTATCA Dog
GTTTCAGAGGCCCCCTCCCCAGGAAACTCCGCAATGACTCACCAGGTATCA Elephant
GTTTCAGAGG------TCCACAGGAACCTTCGCAATGACTTACCAGGTACCA Opossum
GTTTCAGAGGGAC------ACAGGAACCTTCGCA-TGACTAAGCAGGTACCA Chicken
X_tropicalis
Zebrafish
SRSF1

Supplementary Figure 28. Somatic mutations found in COAD and BRCA samples on the alternative exon of MYH11 involved in the isoform switch. The frameshift insertions and deletions fall in a region of low conservation. Right next to these indels we find a putative SRSF1 binding site, with score 7.901 and p-value = 0.000903425. The binding site was predicted using the the program FIMO from the MEME suite and the motif matrix from the RNAcompete datasets.



Supplementary Figure 29. (A) Ranking of p-values for the Mann-Whitney test comparing the PSI difference of tumor and normal isoforms in the switch between the mutated and non-mutated tumor samples. After multiple testing, none of these cases are significant with p-value < 0.05. **(B)** Distribution of the PSI difference of tumor and normal isoforms for the switch in FBLN2 and EHBP1 in BRCA and LUAD, respectively, for mutated and non-mutated tumor samples. A difference can be appreciated, but the number of mutated samples is too small to make a reliable comparison.



(B)

FLNA (COAD)	No mut	mut
No switch	38	5
Switch	208	11
TNC (LUAD)	No mut	mut
No switch	280	9
Switch	188	11
TNC (LUSC)	No mut	mut
No switch	250	8
Switch	221	4
GDF10 (LUAD)	No mut	mut
No switch	179	8
Switch	293	8

Supplementary Figure 30. (A) Relation between mutual-exclusion score and frequency of mutated samples. The mutual-exclusion score measures the proportion of samples that have no switch and have a mutation affecting the protein coding sequence (missense, nonsense, frameshift and indel). **(B)** Number of samples with or without the switch and with (mut) or without (No mut) mutations affecting the protein coding sequence for some of the top ranking cases from (A). These top ranking cases show 9 or less samples for which there are mutations but no switch, but many more (188-293) for which there is a switch but no mutation. **(C)** Ranking of z-scores for the mutual-exclusion scores. Z-scores are calculated as for other measures by comparing to the mutual-exclusion score for 100 random isoforms with similar transcript (total exonic) length using mutations from the same cancer type. The top z-scores occur in switches in cancer types with very few mmutations and are due to just one or two cases. Accordingly, we did not consider the z-score as a relevant measure to assess the occurrence of mutations and switches in a mutually exclusive way.

Supplementary Methods

RNA-Seq data analysis

Available processed RNA-Seq data for tumor and normal samples was downloaded from the TCGA data portal (https://tcga-data.nci.nih.gov/tcga/) for breast carcinoma (BRCA), colon adenocarcinoma (COAD) head and neck squamous cell carcinoma (HNSC), kidney cromophobe carcinoma (KICH), kidney renal clear cell carcinoma (KIRC), lung adenocarcinoma (LUAD), lung squamous cell carcinoma (LUSC), prostate adenocarcinoma (PRAD) and thyroid carcinoma (THCA). In all cases version 2 and Level 3 of the data was used. We used the hg19 version of the annotation from June 2011.

Download locations:

(1)/brca/cgcc/unc.edu/illuminahiseq_rnaseqv2/rnaseqv2/unc.edu_BRCA.IlluminaHiSeq_RNASeqV2.Level_3.1.7.0.tar.gz
(1)/brca/cgcc/unc.edu/illuminahiseq_rnaseqv2/rnaseqv2/unc.edu_BRCA.IlluminaHiSeq_RNASeqV2.mage-tab.1.8.0.tar.gz
(1)/coad/cgcc/unc.edu/illuminahiseq_rnaseqv2/rnaseqv2/unc.edu_COAD.IlluminaHiSeq_RNASeqV2.Level_3.1.7.0.tar.gz
(1)/coad/cgcc/unc.edu/illuminahiseq_rnaseqv2/rnaseqv2/unc.edu_COAD.IlluminaHiSeq_RNASeqV2.mage-tab.1.7.0.tar.gz
(1)/hnsc/cgcc/unc.edu/illuminahiseq_rnaseqv2/rnaseqv2/unc.edu_HNSC.IlluminaHiSeq_RNASeqV2.Level_3.1.5.0.tar.gz
(1)/hnsc/cgcc/unc.edu/illuminahiseq_rnaseqv2/rnaseqv2/unc.edu_HNSC.IlluminaHiSeq_RNASeqV2.mage-tab.1.6.0.tar.gz
(1)/kich/cgcc/unc.edu/illuminahiseq_rnaseqv2/rnaseqv2/unc.edu_KICH.IlluminaHiSeq_RNASeqV2.mage-tab.1.2.0.tar.gz
(1)/kich/cgcc/unc.edu/illuminahiseq_rnaseqv2/rnaseqv2/unc.edu_KICH.IlluminaHiSeq_RNASeqV2.Level_3.1.1.0.tar.gz
(1)/kirc/cgcc/unc.edu/illuminahiseq_rnaseqv2/rnaseqv2/unc.edu_KIRC.IlluminaHiSeq_RNASeqV2.mage-tab.1.4.0.tar.gz
(1)/kirc/cgcc/unc.edu/illuminahiseq_rnaseqv2/rnaseqv2/unc.edu_KIRC.IlluminaHiSeq_RNASeqV2.Level_3.1.3.0.tar.gz
(1)/kirp/cgcc/unc.edu/illuminahiseq_rnaseqv2/rnaseqv2/unc.edu_KIRP.IlluminaHiSeq_RNASeqV2.mage-tab.1.11.0.tar.gz
(1)/luad/cgcc/unc.edu/illuminahiseq_rnaseqv2/rnaseqv2/unc.edu_LUAD.IlluminaHiSeq_RNASeqV2.Level_3.1.12.0.tar.gz
(1)/luad/cgcc/unc.edu/illuminahiseq_rnaseqv2/rnaseqv2/unc.edu_LUAD.IlluminaHiSeq_RNASeqV2.mage-tab.1.13.0.tar.gz
(1)/lusc/cgcc/unc.edu/illuminahiseq_rnaseqv2/rnaseqv2/unc.edu_LUSC.IlluminaHiSeq_RNASeqV2.mage-tab.1.9.0.tar.gz
(1)/lusc/cgcc/unc.edu/illuminahiseq_rnaseqv2/rnaseqv2/unc.edu_LUSC.IlluminaHiSeq_RNASeqV2.Level_3.1.7.0.tar.gz
(1)/prad/cgcc/unc.edu/illuminahiseq_rnaseqv2/rnaseqv2/unc.edu_PRAD.IlluminaHiSeq_RNASeqV2.mage-tab.1.10.0.tar.gz
(1)/prad/cgcc/unc.edu/illuminahiseq_rnaseqv2/rnaseqv2/unc.edu_PRAD.IlluminaHiSeq_RNASeqV2.Level_3.1.8.0.tar.gz
(1)/thca/cgcc/unc.edu/illuminahiseq_rnaseqv2/rnaseqv2/unc.edu_THCA.IlluminaHiSeq_RNASeqV2.mage-tab.1.12.0.tar.gz
(1)/thca/cgcc/unc.edu/illuminahiseq_rnaseqv2/rnaseqv2/unc.edu_THCA.IlluminaHiSeq_RNASeqV2.Level_3.1.11.0.tar.gz

where

(1) = https://tcga-data.nci.nih.gov/tcgafiles/ftp_auth/distro_ftpusers/anonymous/tumor

To assess sample quality, TCGA-provided estimated read-counts per gene were analysed for outliers using URSA (Lee et al. 2013). Using predicted tissue type with posterior P>0.1 we removed samples from the paired dataset that did not cluster with the rest of the samples from the same class (either normal or tumor) using hierarchical clustering with binary distance (presence or absence of a predicted tissue type in a sample) and Ward clustering method (Supplementary Figure 1). We also analysed the non-paired tumor samples, but did not remove any of the samples.

For isoform expression, the abundance of every isoform in each sample was calculated in terms of transcripts per million (TPM) (Li et al. 2010) from the isoform-estimated counts provided by TCGA and the isoform lengths from their annotation (UCSC genes – June 2011). No further normalization on the TPM values was performed. For each isoform *i* in gene G, the PSI was calculated as

$$PSI_i = \frac{TPM_i}{\sum_{\forall k \in G} TPM_k},$$

Genes with one single isoform, or no HUGO ID, were removed for the Iso-kTSP analysis.

The Iso-kTSP algorithm

Iso-kTSP is based on the relative expression reversals of isoforms. It implements the principles of the kTSP algorithm (Geman et al. 2004, Tan et al. 2005) on isoforms from the same gene, with a different scoring function (see below). It reads the TPM values for isoform expression and it stores the rankings in each sample from two possible classes, C_m, $m=1,2$. For every pair of isoforms $I_{g,i}$ and $I_{g,j}$ in each gene g, Iso-kTSP calculates a score based on the frequencies of the two possible relative orders in both classes:

$$S_1(I_{g,i}, I_{g,j}) = P(I_{g,i} > I_{g,j} \mid C_1) + P(I_{g,i} < I_{g,j} \mid C_2) - 1$$

Where $P(I_{g,i} > I_{g,j} | C_1)$ and $P(I_{g,i} < I_{g,j} | C_2)$ are the frequencies at which the isoform $I_{g,i}$ appears later than, or before, $I_{g,j}$ in the expression ranking of class C_1 or C_2, respectively. Score S_1 provides an estimate of how probable are the two isoforms to change relative order between the two classes. Accordingly, the higher the S_1 score, the more consistent the isoform change between classes. To avoid possible ties, a second score, S_2 is used based on the average rank difference per class C_m for each isoform pair, as defined previously (Tan et al. 2005). Defining $R(I_{g,i} | S_a, C_m)$ as the rank of isoform $I_{g,i}$ in sample S_a and class C_m, the average rank difference between two isoforms in a given class is calculated as

$$g(I_{g,i}, I_{g,j} | C_m) = \frac{1}{|C_m|} \sum_a \left(R(I_{g,i} | S_a, C_m) - R(I_{g,j} | S_a, C_m) \right)$$

where $|C_m|$ denotes the number of samples in class C_m. The score S_2 for an isoform pair is then defined as previously (Tan et al. 2005):

$$S_2(I_{g,i}, I_{g,j}) = \left| g(I_{g,i}, I_{g,j} | C_1) - g(I_{g,i}, I_{g,j} | C_2) \right|$$

S_2 provides an estimate of the magnitude of the expression reversal for a pair of isoforms from the same gene between the two classes. All possible isoform pairs are then sorted by the S_1 score and in the case of a tie, by the S_2 score. Even though the isoform expression ranking is global, only pairs of isoforms from the same gene are considered. Additionally, only two isoforms for each gene are chosen, hence a gene can only be listed once in the ranking of isoform pairs.

Classification rules are given in terms of *k* isoform pairs. The classification of a new sample is performed by evaluating each isoform-pair rule against the ranking of isoform expression of the new sample. For each isoform-pair rule, the classifier selects the class for which the data fulfills the rule. The final decision for classification is established by simple majority voting, by selecting the most voted class from the *k* rules. In order to avoid ties in this voting, *k* is always odd.

The optimal number *k* of isoform pairs in the classifier, k_{opt}, is calculated by performing cross-validation on the training set. The samples are split into training and testing sets, each containing the same number of samples from each class. For our analyses, at each step of the cross-validation the testing was done at a single pair of tumor-normal, training on the rest of the paired samples; hence, at the end of the cross-validation every tumor-normal pair was tested once. At each step of the cross-validation, a model is built for various values of *k* odd, up to a value k_{max} input as parameter (k_{max} = 50 in our analysis). For each *k*, the accuracy of the model is evaluated against the test set, where

$$\text{accuracy} = \frac{TP + TN}{TP + TN + FN + FP}$$

and where TP, TN, FN and FP are the true positives, true negatives, false negatives and false positives, respectively. This accuracy value is symmetric with respect to the choice of either class as reference for positive cases. After cross-validation, the optimal number of pairs, k_{opt}, is taken to be the minimum k with the highest average accuracy, where the average is calculated across all cross-validation steps. Finally, using the global ranking of isoform pairs, now calculated on the entire dataset, the top k_{opt} pairs are then reported as the classification model. These k_{opt} pairs are the isoforms that most consistently and with greatest difference change relative expression between two classes (e.g. tumor and normal, two cancer subtypes, etc.) and that best separate these two classes. Additionally, Iso-kTSP also reports the individual performances of each individual isoform pair in terms of the Information Gain measure, which provides an estimate of the discriminating power of each individual isoform pair, and which can be written in terms of the TP, TN, FN and FP values for each individual isoform pair rule. For a balanced set:

$$IG = 1 - PPV \cdot \log_2 PPV - FDR \cdot \log_2 FDR - NPV \cdot \log_2 NPV - FNR \cdot \log_2 FNR$$

where PPV, FDR, NPV and FNR are the positive predictive value, false discovery rate, negative predictive value and false negative rate, respectively:

$$PPV = \frac{TP}{TP + FP}, \; FDR = \frac{FP}{TP + FP}$$
$$NPV = \frac{TN}{TN + FN}, \; FNR = \frac{FN}{TN + FN}$$

The algorithm Iso-kTSP is implemented in Java. Software and documentation are available at https://bitbucket.org/regulatorygenomicsupf/iso-ktsp

More information is provided at https://bitbucket.org/regulatorygenomicsupf/iso-ktsp/wiki/Home

Z-score calculation for Score S1 and Information Gain

To calculate a Z-score for Score S1 and the information gain (IG), we carried out an Iso-kTSP analysis, where we calculated the single pair performance for all genes used in the analysis. That is, asking the algorithm to provide a ranking of isoform pairs for all genes, which provides the best isoform-pair per gene, with the corresponding individual IG and S_I values. This resulted in 14630 values for S_I and IG, which were used to calculate mean values and standard deviation of the population. We then subtracted the mean from each individual value and divided it by the standard deviation, resulting in a standardized Z-score.

Permutation analysis

Significance of the computed isoform switches was evaluated with Iso-kTSP program using permutation analysis. Namely, labels from the two classes (e.g. tumor vs normal) were shuffled 1000 times. For each shuffling step, the algorithm described above was re-run and the top-scoring isoform-pair was selected. The distribution of the 1000 top-scoring pairs from the shufflings provides an estimate of the expected recurrence of isoform changes obtained by chance (Supplementary Figure 10). An isoform-pair is significant if its Information Gain and Score S_I are both larger than any of the values obtained from the 1000 shufflings.

Blind tests

A blind test of each model was carried out on the set of samples that were not used for cross-validation. These are all the unpaired tumor samples. For the analysis of the cancer subtypes, however, the overall accuracy is estimated on the complete set of tumor samples. In all these tests, we measure the accuracy as described above, calculating the proportion of samples correctly labeled by the classifier, as well as counting the number of correct votes per sample.

Anticorrelation

For each isoform pair in a rule we calculated the Spearman correlation of the PSI values for the paired tumor and normal samples. Those pairs with correlation value $R < -0.8$, i.e. anticorrelation, were considered proper isoform switches where the most abundant isoform in each sample is one of the isoforms in the pair.

Analysis of Subtypes

We applied the Iso-kTSP algorithm to study the four BRCA subtypes as classified by TCGA (TCGA 2012): luminal A, luminal B, basal and Her2+, which have 231, 127, 98 and 58 samples, respectively. For each subtype, 100 random subsets were sampled, selecting 45 from a given subtype, and 15 from each of the other subtypes to maintain balanced sets. For each one of the 100 iterations, an Iso-kTSP model was built as described above with k = 50; and the significance of the predicted isoform changes were computed using permutation analysis as described above. Finally, for each subtype and for each isoform pair, the proportion of iterations in which it is used in a model and the proportion of iterations in which it is significant is reported.

We also analysed the four LUSC subtypes: basal, classical, primitive and secretory (Wilkerson et al. 2010), with 41, 57, 25 and 38 samples, respectively. Similarly to the BRCA subtypes, 100 random samples for each subtype were considered, each time selecting 24 from one subtype and 8 from each other three subtypes. For each of the 100 iterations Iso-kTSP model was built with k = 50, and the significance estimated by permutation analysis.

COAD non-paired tumor samples were separated into hypermutated or non-hypermutated according to the number of mutations per sample. Using all mutation types (classified by TCGA as functional or non-functional), samples with more than 250 mutations were classified as hypermutated and those with less than 250 mutations were classified as non-hypermutated; leaving 59 and 130 samples, respectively. Similarly as before, both COAD subtypes were compared by sampling 100 times 40 samples from either subtype, and by building the Iso-kTSP model at each iteration and computing the significance by permutation analysis.

Analysis of somatic mutations

We downloaded somatic mutation data from TCGA. Download locations:

(1)/brca/gsc/genome.wustl.edu/illuminaga_dnaseq_curated/mutations/genome.wustl.edu_BRCA.IlluminaGA_DNASeq_curated.Level_2.1.1.0.tar.gz
(1)/brca/gsc/genome.wustl.edu/illuminaga_dnaseq/mutations/genome.wustl.edu_BRCA.IlluminaGA_DNASeq.Level_2.5.3.0.tar.gz
(1)/coad/gsc/hgsc.bcm.edu/solid_dnaseq/mutations/hgsc.bcm.edu_COAD.SOLiD_DNASeq.Level_2.1.7.0.tar.gz
(1)/coad/gsc/hgsc.bcm.edu/illuminaga_dnaseq/mutations/hgsc.bcm.edu_COAD.IlluminaGA_DNASeq.Level_2.1.5.0.tar.gz
(1)/hnsc/gsc/broad.mit.edu/illuminaga_dnaseq/mutations/broad.mit.edu_HNSC.IlluminaGA_DNASeq.Level_2.1.0.0.tar.gz
(1)/kich/gsc/hgsc.bcm.edu/mixed_dnaseq_curated/mutations/hgsc.bcm.edu_KICH.Mixed_DNASeq_curated.Level_2.1.0.0.tar.gz
(1)/kich/gsc/broad.mit.edu/illuminaga_dnaseq/mutations/broad.mit.edu_KICH.IlluminaGA_DNASeq.aux.1.0.0.tar.gz

(1)/kich/gsc/broad.mit.edu/illuminaga_dnaseq/mutations/broad.mit.edu_KICH.IlluminaGA_DNASeq.Level_2.1.0.0.tar.gz
(1)/kich/gsc/hgsc.bcm.edu/illuminaga_dnaseq/mutations/hgsc.bcm.edu_KICH.IlluminaGA_DNASeq.Level_2.1.2.0.tar.gz
(1)/kirc/gsc/hgsc.bcm.edu/mixed_dnaseq/mutations/hgsc.bcm.edu_KIRC.Mixed_DNASeq.Level_2.1.2.0.tar.gz
(1)/kirc/gsc/broad.mit.edu/illuminaga_dnaseq/mutations/broad.mit.edu_KIRC.IlluminaGA_DNASeq.Level_2.1.5.0.tar.gz
(1)/kirc/gsc/broad.mit.edu/illuminaga_dnaseq/mutations/broad.mit.edu_KIRC.IlluminaGA_DNASeq.aux.1.2.0.tar.gz
(1)/luad/gsc/broad.mit.edu/illuminaga_dnaseq/mutations/broad.mit.edu_LUAD.IlluminaGA_DNASeq.Level_2.0.4.0.tar.gz
(1)/lusc/gsc/broad.mit.edu/illuminaga_dnaseq/mutations/broad.mit.edu_LUSC.IlluminaGA_DNASeq.Level_2.100.1.0.tar.gz
(1)/lusc/gsc/broad.mit.edu/illuminaga_dnaseq/mutations/broad.mit.edu_LUSC.IlluminaGA_DNASeq.aux.1.2.0.tar.gz
(1)/prad/gsc/hgsc.bcm.edu/illuminaga_dnaseq_automated/mutations/hgsc.bcm.edu_PRAD.IlluminaGA_DNASeq_automated.Level_2.1.0.0.tar.gz
(1)/prad/gsc/broad.mit.edu/illuminaga_dnaseq_curated/mutations/broad.mit.edu_PRAD.IlluminaGA_DNASeq_curated.Level_2.1.2.0.tar.gz
(1)/prad/gsc/broad.mit.edu/illuminaga_dnaseq/mutations/broad.mit.edu_PRAD.IlluminaGA_DNASeq.Level_2.1.3.0.tar.gz
(1)/prad/gsc/broad.mit.edu/illuminaga_dnaseq/mutations/broad.mit.edu_PRAD.IlluminaGA_DNASeq.aux.1.0.0.tar.gz
(1)/thca/gsc/hgsc.bcm.edu/illuminaga_dnaseq_automated/mutations/hgsc.bcm.edu_THCA.IlluminaGA_DNASeq_automated.Level_2.1.1.0.tar.gz
(1)/thca/gsc/broad.mit.edu/illuminaga_dnaseq/mutations/broad.mit.edu_THCA.IlluminaGA_DNASeq.aux.1.5.0.tar.gz
(1)/thca/gsc/broad.mit.edu/illuminaga_dnaseq/mutations/broad.mit.edu_THCA.IlluminaGA_DNASeq.Level_2.1.5.0.tar.gz

where, as before,

(1) = https://tcga-data.nci.nih.gov/tcgafiles/ftp_auth/distro_ftpusers/anonymous/tumor

The mutations analysed include all mutations as annotated by TCGA, regardless of the functional categorization. For the purpose of finding associations, we simply consider whether a gene or the isoform-pair of the gene in a given sample either has or does not have mutations. Although most of the mutational data in TCGA comes from exome sequencing data, there are cases where it comes from whole genome sequencing. Accordingly, in order to homogenize the mutation sampling, we considered mutations that fall on exon regions or in the 100 nt upstream or downstream flanking regions. We carried out several different tests for the association of somatic mutations and isoform switches.

Jaccard index

Given the samples with one or more mutations *M* on the transcripts involved in the switch, and the samples with the isoform switch *S*, a Jaccard index *J* for the association of these two variables was calculated:

$$J = \frac{|M \cap S|}{|M \cup S|}$$

This index was calculated using the clujaccard function from the fpc package of R (http://cran.r-project.org/web/packages/fpc). A Z-score from the Jaccard index was calculated by comparing to genes with similar median isoform length. Genes used in the Iso-kTSP analysis were ranked according to the absolute length difference of the median isoform length with the given gene. Lengths were calculated as the sum of the component exons from the TCGA annotation (UCSC genes, June 2011), which is equivalent to the length of the projection of the exons from the gene to the genome. The top 100 with the smallest length difference were selected, and two of its isoforms were chosen randomly. For each of these matched controls the Jaccard index was calculated as before. A Z-score was then calculated based on the mean and standard deviation of the Jaccard index from the 100 matched controls. The above analysis was also repeated using only mutations that fall on exons and that affect the protein sequence (Frame_Shift_Del, Frame_Shift_Ins, In_Frame_Del, In_Frame_Ins, Missense_Mutation, Nonsense_Mutation, Nonstop_Mutation) with very similar results.

Wilcox-test of delta-PSI values

For each isoform-pair we calculated a delta-PSI value for each tumor sample by subtracting the PSI of the tumor isoform from the PSI of the normal isoform. We then separated the samples into two classes, based on the presence or absence of any mutation in the isoform pair. We carried out a Mann-Whitney test to check if there are significant differences in the distribution of delta-PSI values between the two populations, and used the Benjamini-Hochberg method to correct the p-values for multiple testing. The same analysis was also performed using mutations on the entire gene locus with very similar results.

Fisher test

A Fisher test was performed per isoform-pair by building a 2x2 contingency table with the number of tumor samples with or without mutations and the number of tumor samples with or without the corresponding switch. None of the isoform switches had a significant association according to this test after correcting for multiple testing with the Benjamini-Hochberg method. The same analysis was also performed using mutations on the entire gene locus with very similar results.

Mutual information

The mutual information was calculated using the presence or absence of mutations and switches in each sample, using the mi.empirical function from the entropy package of R (Hausser et al. 2009). Using mutations falling on the isoforms undergoing the switch, the mutual information values correlate with the Jaccard index (Spearman R = 0.9) and with the Jaccard z-scores (Spearman R=0.7) calculated above. This analysis was also performed using mutations falling on the entire gene locus with very similar results.

Mutual exclusion analysis of functional mutations and switches

A mutual-exclusion between isoform switches and protein-affecting mutations was measured as follows: given the number of samples having an isoform switch and no mutation (n_{10}), and those having a mutation but no isoform switch (n_{01}), a mutual-exclusion score was defined to be:

$$2\frac{\min(n_{10}, n_{01})}{N}$$

where N is the total number of samples. A z-score was calculated similarly as above for the Jaccard index.

Reactome analysis

For the Reactome (Croft et al. 2014) enrichment analysis we used the ReactomePA package from Bioconductor (Yu 2014) with default values, only changing the "universe" parameter, where we used the 14630 Entrez gene IDs that were used for the Iso-kTSP analysis. The ReactomePA

GO enrichment analysis

GO-terms were tested for enrichment in the three ontologies: Biological Process, Cellular Component and Molecular Function. The GOstats package (Falcon and Gentleman 2007) was used for the test, with a P-value cutoff of 0.05, odds-ratio > 2 and at least 5 genes. The background set was the gene set used for the Iso-kTSP analysis and the "conditional" option was used in the test to calculate the significance of a parent term, if its children were significant.

Supplementary Files description:

Supplementary File 1 (SupplementaryFile1.tgz)

This is a compressed file with the lists of patient-samples used for the analyses. Each sample is described by a short ID, which was also used in the data tables. The sample ID is the participant ID part of the TCGA barcode, with an additional N or T character at the end, meaning normal or tumor sample, respectively.

Supplementary File 2 (SupplementaryFile2.tgz)

This is a compressed file with the Iso-kTSP models ready to be used and tested with the Iso-kTSP algorithm. The format is plain text, where each isoform-pair of the model is given per line, separated by a tab. Each isoform is specified by an ID composed of the gene-ID and transcript-ID separated by “,”, e.g. the LUAD model is:

QKI|9444,uc003qui.2	QKI|9444,uc003qug.2
NUMB|8650,uc001xnz.1	NUMB|8650,uc001xoa.1
MTA3|57504,uc002rsr.2	MTA3|57504,uc002rsp.1
SLC39A8|64116,uc003hwb.1	SLC39A8|64116,uc011ceo.1
ESYT2|57488,uc003wod.1	ESYT2|57488,uc003wob.1
PRMT2|3275,uc002zjy.2	PRMT2|3275,uc002zjz.1
ISLR|3671,uc002axh.1	ISLR|3671,uc002axg.1

Supplementary File 3 (SupplementaryFile3.tar.gz)

This (tar.gz) file includes the GFF (version GFF3) files of the tumor and normal isoforms from the switches and the Iso-kTSP models.

Supplementary File 4 (SupplementaryFile4.txt)

This is a plain text, tab-separated, file with all the information derived for the isoform switches. The file contains a header. The column description of the Supplementary File 4 is as follows:

Header	Explanation

Cancer	Cancer type (BRCA, etc..)
gene_name	HUGO gene ID
tumor	Tumor isoform ID, format: gene_name,transcript_id
Normal	Normal isoform ID, format: gene_name,transcript_id
Ig	Information Gain of the switch
score1	Score S_1 of switch
Ig_zscore	Z-score of switch information gain
score1_zscore	Z-score of switch Score S_1
corr	Spearman correlation of isoform-pair PSI values in the paired samples
TinN	Average expression TPM of tumor isoform in normal samples
TinT	Average expression TPM of tumor isoform in tumor samples
NinN	Average expression TPM of normal isoform in normal samples
NinT	Average expression TPM of normal isoform in tumor samples
Isonum	Total number of annotated isoforms in gene
freqAG	Frequency of samples with any mutation in the full gene
freqAI	Frequency of samples with any mutation in the isoform switch
freqFG	Frequency of samples with functional mutation in the full gene
freqFI	Frequency of samples with functional mutation in the isoform switch
dpsiAG	Mann-Whitney test p-value for comparing PSI values of tumor samples with and without any mutation in the full gene
dpsiAI	Mann-Whitney test p-value for comparing PSI values of tumor samples with and without any mutation in the isoform switch
freq_switches	Frequency of samples with the isoform switch
fisherAG	Fisher test p-value comparing the association of isoform switches and any type of mutations along the full gene in the samples
fisherAI	Fisher test p-value comparing the association of isoform switches and any type of mutations along the isoform switch in the samples
mutinfAG	Mutual information score of the association of isoform switches and any type of mutations along the full gene in the samples
mutinfAI	Mutual information score of the association of isoform switches and any type of mutations along the isoform switch in thea samples
jaccardAG	Jaccard index of the association of isoform switches and any type of mutations along the full gene in the samples
jaccardAI	Jaccard index of the association of isoform switches and any type of mutations along the isoform switch in the samples
jaccardFG	Jaccard index of the association of isoform switches and functional mutations along the full gene in the samples
jaccardFI	Jaccard index of the association of isoform switches and functional mutations along the isoform switch in the samples
funvalFG	Mutual exclusion score of the association of isoform switches and functional mutations along the full gene in the samples
funvalFI	Mutual exclusion score of the association of isoform switches and functional mutations along the isoform switch in the samples
jaccardAG_Z	Z-score of the jaccardAG score
jaccardAI_Z	Z-score of the jaccardAI score
jaccardFG_Z	Z-score of the jacardFG score
jaccardFI_Z	Z-score of the jaccardFI score
funvalFG_Z	Z-score of the funvalFG score
funvalFI_Z	Z-score of the funvalFI score
dpsiAG_adj	Benjamini - Hochberg adjusted dpsiAG p-value
dpsiAI_adj	Benjamini - Hochberg adjusted dpsiAI p-value
tumoriso_transcript	Tumor isoform mRNA length
tumoriso_genomic	Tumor isoform pre-mRNA length
normaliso_transcript	Normal isoform mRNA length
normaliso_genomic	Normal isoform pre-mRNA length

Data availability

Data is available at DOI: http://dx.doi.org/10.6084/m9.figshare.1061917

References

Croft D, Mundo AF, Haw R, Milacic M, Weiser J, Wu G, Caudy M, Garapati P, Gillespie M, Kamdar MR, Jassal B, Jupe S, Matthews L, May B, Palatnik S, Rothfels K, Shamovsky V, Song H, Williams M, Birney E, Hermjakob H, Stein L, D'Eustachio P (2014) The Reactome pathway knowledgebase. Nucleic Acids Research 1 January 2014; 42(D1):D472-D477

Falcon S, Gentleman R. (2007) Using GOstats to test gene lists for GO term association. Bioinformatics 23(2):257-8

Geman D, d'Avignon C, Naiman DQ, Winslow RL. Classifying gene expression profiles from pairwise mRNA comparisons. Stat Appl Genet Mol Biol. 2004;3:Article19.

Hausser J, & Strimmer K. 2009. Entropy inference and the James-Stein estimator, with application to nonlinear gene association networks. J. Mach. Learn. Res. 10: 1469-1484

Lee YS, Krishnan A, Zhu Q, Troyanskaya OG. Ontology-aware classification of tissue and cell-type signals in gene expression profiles across platforms and technologies. Bioinformatics. 2013 Dec 1;29(23):3036-44.

Li B, Ruotti V, Stewart R, Thomson J, Dewey C. (2010). RNA-Seq gene expression estimation with read mapping uncertainty. Bioinformatics 26(4):493-500.

Tan AC, Naiman DQ, Xu L, Winslow RL, Geman D. Simple decision rules for classifying human cancers from gene expression profiles. Bioinformatics. 2005 Oct 15;21(20):3896-904.

The Cancer Genome Atlas Network (2012c) Comprehensive molecular portraits of human breast tumours. *Nature.* 490(7418):61-70

Wilkerson MD, Yin X, Hoadley KA, Liu Y, Hayward MC, Cabanski CR, Muldrew K, Miller CR, Randell SH, Socinski MA, Parsons AM, Funkhouser WK, Lee CB, Roberts PJ, Thorne L, Bernard PS, Perou CM, Hayes DN. Lung squamous cell carcinoma mRNA expression subtypes are reproducible, clinically important, and correspond to normal cell types. Clin Cancer Res. 2010 Oct 1;16(19):4864-75.

Yu G. (2014) *ReactomePA: Reactome Pathway Analysis*. R package version 1.8.1.

http://www.bioconductor.org/packages/release/bioc/html/ReactomePA.html